THE UNIVERSITY
of EDINBURGH

Effects of Andean geographic dynamics on the population history of *Tococa*-associated *Azteca* ants

María Fernanda Torres Jiménez

Doctor of Philosophy

School of Biological Sciences

Institute of Evolutionary Biology

University of Edinburgh

2017

Table of Contents

Abstract

Myrmecophytic plant species form associations where the ant colony inhabits structures in the plant and offers protection against herbivory in exchange for food and shelter. Widely distributed across the tropics, myrmecophytic mutualisms are particularly diverse in the Neotropics, a region characterized by the rapid and recent uplift of the Andean mountain range. It has been suggested that the abrupt change in terrain triggered the emergence of new niches, new barriers to gene flow and speciation. Studying ant-plant associations in the Neotropics not only provides insight into how associations evolve in time but also the impact that external factors, such as geographic changes, have in the evolution of mutualisms.

Because of its wide distribution on both sides of the Andes, The *Tococa guianensis-Azteca* system is useful to explore the effects the Andean uplift had on the evolution of mutualisms. This thesis aims to *1.* Identify the ants associating with *T. guianensis* and the lineages of ants and plants involved in the mutualisms in different populations on both sides of the Andes, *2.* generate genomic data for both ants and plants to increase sampling of loci, and *3.* estimate and calibrate the species trees to compare patterns of phylogenetics and temporal congruence between ants, plants and the Andean uplift. Most ant-plant studies focus on only one partner or study both partners by using already collected data for one of them. This project is the first study inferring the evolutionary history of both partners associated at that point in time and across a large area.

This thesis identifies two main *Azteca* lineages associated with *T. guianensis*, each one distributed on different sides of the Andes. It addresses the monophyly of *T. guianensis* (and related species) and why such monophyly cannot be confirmed. Results show how both plants and ants were geographically structured congruent with timing of a split of populations coinciding with the Andean uplift. Moreover, four plants and fifteen ant genomes were assembled and used to estimate gene and species trees. For *Tococa*, candidate markers were selected for future resolution of the plant's phylogeny.

Different histories but similar divergence times between ants and plants suggest that the mutualism has evolved in response to geographic changes rather than through codiversification, but that the mutualism persists thanks to the availability of the host. The information generated during this study provides the basis to understand the evolution of mutualisms, the genomic features of ants and plants and opens the possibility for *Tococa* and *Azteca* to become a model system.

Lay Summary

Some ant and plant species form associations where the ant colony inhabits structures in the plant and offers protection against herbivory in exchange for food and shelter. Widely distributed across the tropics, ant-plant mutualisms are particularly diverse in the Neotropics, a region characterized by the rapid and recent rising of the Andean mountain range. Abrupt changes in terrain triggered the appearance of new environmental conditions, limited the exchange of genetic material and fostered the emergence of new species. Studying ant-pant associations in the Neotropics not only provides an insight on how associations evolve in time but also the impact that external factors, such as geographic changes, have in the evolution of mutualisms.

Because of its wide distribution on both sides of the Andes, The *Tococa guianensis-Azteca* system is useful to explore the effects the mountain range had in the evolution of mutualisms. This thesis aims to *1.* Identify the ants inhabiting *T. guianensis* in the different populations on both sides of the Andes, *2.* generate genomic data for both ants and plants to increase the data available for analyses, and *3.* estimate the time of events when populations were isolated and compare this with the times of the Andean uplift. This project is the first study inferring the evolutionary history of both ant and plant individuals associated at a single point in time and across a large area.

This thesis identifies two *Azteca* groups associated with *T. guianensis*, each one distributed on different sides of the Andes. Results show how both plants and ants were geographically structured congruently with the presence of the Andes cordillera. Moreover, four plants and fifteen ant genomes were sequenced and used to estimate and compare plant and ant evolutionary histories.

Different histories but similar divergence times between ants and plants suggest that the mutualism has evolved in response to geographic changes. The information generated during this study provides the basis to understand the evolution of mutualisms, the genomic features of ants and plants and opens the possibility for *Tococa* and *Azteca* to become a model system.

Acknowledgements

In no particular order:

To my PhD supervisors Doctors James E. Richardson and Graham Stone who provided academic and life-saving advise, sometimes in the generous shape of distilled sugar cane juice. To my fieldwork partner and sofa-provider Julieth Serrano who suffered with me the inclement Colombian transportation system, strikes, heat, dehydration and one or two retrospectively funny situations we managed to get out from. To Flavia Pezzini for the help and company in the RBGE lab. To Karina Banda, Eugenio Valderrama, Maca Gomez and Andres Orejuela, who with Julieth and I were the imported Colombian team taking over control of Edinburgh (and thanks again to James R. who orchestrated the invasion and who is himself an honorary Colombian). To Adriana Sanchez, who inspired me to be a scientist. To James Nicholls (a.k.a. Hames) who was always there to correct me and teach me, and from whom I learnt useful lab hacks. To Maria Pinilla, Juliana Cardona and Javier F. Tabima who sent from Colombia endless support in the shape of gifs and memes. To Tom Godfrey, Nora Villamil, Julja Ernst, Lisa Cooper, Lisa Gecchele, Luiz Carvalho, Andres de la Filia, Doris Reineke, Hansi, and Jack Shutt who were loyal pub company in the hard moments. To Sujai Kumar for bioinformatic help and friendship, and to Jack Hearn for all the help with some of the methods. To Gytis Dudas: best beer provider, Secret Aardvark smuggler, peloton member, company at all times, support under all moods and T.V. series partner. To my family for all the unconditional support and love. To my brother, Londonian franchise of my family and biggest support in all times. To my cat, who knew me the most. To all fieldwork assistants who, no matter the situation, did not run away and let me helpless in the middle of the jungle (except for those who did). To all Scottish breweries, Sainsbury's super-hot spicy sauce and Secret Aardvark for providing all the energy required during this process. To Hector Lavoe, Piper Pimienta and Willie Colon, loyal friends. This has been a great experience. ♪

To the Darwin Trust of Edinburgh for funding my Doctorate Degree and to the Davis Fund for fieldwork funding.

Thanks to all journals for the permissions to reproduce figures in the manuscript of this thesis.

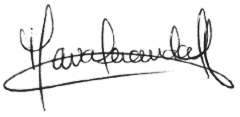
Plants were collected under the permit **0738 from July 8th, 2014** granted to the Universidad Distrital, Colombia. Fertile plant vouchers representative of each population are at the Herbario Forestal UDBC. Voucher numbers will be provided when the exicates are formally entered.

Ants were collected under the permit **0530 from May 27th, 2014** granted to the Universidad del Rosario, Colombia. Ants are preserved in 96% Ethanol in cryovials at -80°C at the Colección de Invertebrados de la Universidad del Rosario. Voucher numbers will be provided after fulfilling with a final report to the government institution granting the permit ANLA (National Authority for Environmental Licensing) in December 2018.

Declaration

I declare that this thesis was composed by myself and that all the work described within is my own. It has not been submitted for any other degree of professional qualification except as specified.

María Fernanda Torres Jiménez, 2017



CHAPTER

1

INTRODUCTION

1.1 Generalities of mutualisms

Mutualisms are positive interactions between individuals from different species in which the net effect for both is beneficial. It involves the exchange of goods produced by a provider that the recipient cannot obtain on its own or its production is expensive for the recipient (Bronstein, 1994; Schwartz and Hoeksema, 1998; Herre et al., 1999; Leigh, 2010). Examples of mutualisms are common in nature: plants produce sugar-rich fruits to favor dispersal by monkeys, fig wasps reproducing inside fig syconia while pollinating the flowers, nitrogen-fixing cyanobacteria living inside corals, and ants protecting and feeding on plants they inhabit. In these examples, the energy invested by one mutualist

in the other is paid back by the benefits obtained from the other. Mutualist partners interact at different degrees of dependency, from facultative associations to obligate ones, in which partners are highly specialized (Douglas, 2010; Leigh, 2010). It is possible that cheaters appear in the mutualism, and in these cases of parasitism, evolution selects for mechanisms to stabilize the interaction and punish the cheaters.

Mutualisms are established each generation by either vertical or horizontal transmission, each one with implications for the strength, stability and evolution of the association. Vertical transmission is common between partners in a symbiosis (Box 1.1), where new generations of hosts inherit symbionts directly from their parents and the symbiont spends its life cycle inside the host. A common example is the maternal transmission of a beneficial strain of *Wolbachia* (wRi) in populations of *Drosophila simulans*. The infection is passed from females to their offspring because the bacteria infect the eggs (Werren, 1997; Weeks et al., 2007). Similarly, some fungi-growing Attini ant queens carry a fungal cultivar from their colony of origin to the one the queen is about to found; however, horizontal transmission of the fungi is possible (Weber, 1966; Mueller, 2002). In the other hand, horizontal transmission requires that every new generation of mutualists identify and find each other, and the life cycle of one partner does not take place entirely inside the other. Ant-plant mutualisms are associations transmitted horizontally when the reproductive alates mate outside the plants and search for a new host (Davidson and McKey, 1993; Bruna et al., 2011). When comparing patterns of evolution between partners, vertically transmitted symbionts are expected to produce concordant phylogenies with those of their hosts, while incongruities and switches are expected from horizontal transmission (Herre et al., 1999). These expectations are not always the rule and later I will discuss how different processes lead to similar patterns and vice versa.

Box 1.1. Glossary

Coadaptation
Is the array of two or more entities (species, populations, genes, traits) that exert reciprocal selection pressures to minimize possible disadvantageous effects (McFarquhar and Robertson, 1963; Wade, 2007). Coadaptation can occur with or without coevolution (Janz, 2011). Coevolution will lead to coadaptation when the response to reciprocal selection pressures is inherited and evolved together (see the definition of coevolution below) (Wade, 2007). But coadapted entities do not always arise by coevolution. In this case, coadapted traits appear in non-interacting ancestors as a response to different selective pressures and become coadapted as the populations bearing the traits come into contact (Ridley, 2003)

Coevolution
Coevolution refers to the reciprocal selecting pressure exerted by two or more different lineages of closely interacting organisms and that results in evolutionary changes in traits involved (directly or indirectly) in the association (Ehrlich and Raven 1964; Thompson 1994; Dale H. Clayton 1997; Segraves 2010; Althoff et al. 2014). It was first defined by Ehrlich and Raven (1964) as the reciprocal evolution of organisms in response to one another in a process that occurs in a stepwise manner. Similarly, Janzen (1980) defines coevolution as the change in traits present in one population in response to changes in traits of another population, followed by a subsequent evolutionary change in the second population in response to the first. Coevolution can be pairwise or diffuse, depending on the specificity of the association, their symmetry, and the number of populations involved (Janzen, 1980; Fox, 1988; Janz, 2011). *Pairwise coevolution* is expected to occur in one-to-one specific interactions and if evolutionary responses between the two species have no impact on their interactions with other species (Janzen, 1980; Futuyma, 2009). In less specific interactions when one or both populations are arrays of populations, *diffuse coevolution* is expected (Janzen, 1980). In this case, changes in one population can affect one or more populations (Futuyma, 2009).

Codiversification
Codiversification is the degree of correlation between divergence events occurring in two or more lineages of organisms (de Vienne et al., 2013; Althoff et al., 2014), and requires that divergence events happen at similar absolute times (Herre et al. 1999, see B and C in Figure 1.1). Codiversification can emerge from coevolutionary processes or for shared histories between organisms. For instance, codiversification it can emerge because of shared geographic, ecological and climate conditions causing co-distributed populations of different organisms to split (Herre et al., 1999; Segraves, 2010; Hembry et al., 2014). Cospeciation processes occurring at a lower level than species (Page, 2003). Synonym terms in the literature are cocladogenesis or parallel cladogenesis.

Cospeciation
Congruent evolutionary histories producing matching phylogenies and phylogenetic differentiation between interacting species (Hafner and Nadler, 1990; Futuyma, 2009). It can result from the interactions, from shared geographic histories or a mixture of both (Futuyma, 2009).

Host switch
Or host shift. Occurs when a lineage changes the host to which it normally associates with. Often, incongruence between host and symbiont phylogenies is evidence of host switching (Page, 1993); however, a switch between closely related host lineages can result in congruent phylogenies (Vienne et al., 2007; Janz, 2011).

Mutualism
Interaction in which populations or species use each other as a resource (*i.e.* reciprocal exploitation), and in which the cost of providing the resource is lower than the benefit obtained (Bronstein, 1994; Futuyma, 2009). Mutualisms can occur between free-living organisms (*e.g.* plants and pollinators), or between one organism spending most of its life cycle on or in another organism (*e.g.* legumes and *Rhizobium* bacteria).

Parasitism
Interaction in which a population exploits another as a resource, gaining benefits without providing any and representing a cost to the second population (Page, 1993).

Symbiosis
According to De Bary (1879), symbiosis is an "intimate, outcome-independent interaction between species" were intimate means that one organism lives inside the other. Other authors are more specific and use symbiosis to define beneficial, intimate non-parasitic associations (McNaughton and Wolf, 1973; Saffo, 1992). Both definitions fit the description of ant-plant mutualisms as a mutually beneficial interaction and the use of the term is not ambiguous. However, the term symbiont can pose some ambiguity. Even though plant-ants spend most of their time inside or on the plant, fertile individuals mate outside and the mutualism is horizontally inherited. These ant-plant interactions are not strictly comparable with interactions between bacteria and for instance, insects, were the bacteria completes their life cycle inside the host and the mutualism is vertically inherited. Nevertheless, a large portion of the life cycle of obligate plant-ants takes place inside the host's domatia; thus, the term symbiont will be used throughout the text to refer to the organisms inhabiting the host in mutualistic interactions.

1.1.1 Coevolution, codiversification, and the maintenance of diversity

Mutualisms play important ecological roles and are associated with the maintenance of biodiversity (Ehrlich and Raven, 1964; Bascompte and Jordano, 2007; Hembry et al., 2014). For instance, diversification of flowering plants is thought to partially result from insect diversification, which in turn evolved as new niches become available when angiosperms first appeared (Raven, 1977; Regal, 1977; Burger, 1981; Crane et al., 1995; Grimaldi and Engel, 2005). Such diversity is likely promoted by coevolutionary and co-diversification processes between mutualists, such as between plants and pollinators, or animal dispersal of seeds and angiosperms (Jordano, 2000). Mechanisms operating at the population level and genotype by genotype by environmental interactions (referred to as GxGxE by Thompson 2005) can translate into diversification in many ways. Let's imagine a mutualism whose distribution range is large. Reciprocal selection between hosts and symbionts in a race to keep cheaters out of the association can lead to changes in allele frequencies of association-related traits over successive generations. Provided different ecological backgrounds on opposite extremes of the distribution range, reciprocal variation can arise locally and eventually result in new lineages of coevolving organisms. Mutualisms can also increase diversity via codiversification and coevolution in a similar way to the "escape and speciate" model (Ehrlich and Raven, 1964). If host populations are isolated by the appearance of a new barrier or habitat disruption, the reduced gene flow can potentiate lineage divergence and ultimately speciation. Even if gene flow between symbiont populations is not affected, genetic drift can influence allele frequencies of traits related to, for instance, symbiont choice and end up boosting the divergence between different symbiont populations. Finally, mutualisms can promote diversification when a host undergoes independent speciation (due to vicariance

or divergent selection) and the new host lineages become niches that phylogenetically unrelated symbionts can exploit and potentially speciate as a result.

No organism is completely isolated from its environment and neither are the other organisms cohabiting it. As much as environmental and ecological changes can determine the output of evolutionary processes, associations between different taxa can shape their evolutionary path. The different outcomes leading to an increase of diversity in mutualisms can be explained by the different combinations in the magnitude of coevolutionary and codiversification processes and the strength of the reciprocal pressures on traits mediating the association. Different definitions of coevolution and codiversification exist in literature, and because both are not mutually exclusive mechanisms, it is easy to assume that they are highly correlated, but that is not always the case (Janzen 1980; Janz 2011 and see Box 1.1). Testing for coevolution requires a clear assessment of correlations between traits to determine that a change has been in fact promoted by the interactions with the associated lineage. On the other hand, codiversification or cocladogenesis requires the parallel divergence of different taxa to happen at similar absolute times (Herre et al. 1999 and can be the result of abiotic factors not related to coevolutionary processes (see B and C in Figure 1.1). For this thesis, coevolution will be assumed as the reciprocal selection upon traits mediating the association (disregarding whether is pairwise or diffuse) and codiversification will be assumed as the topological and temporal congruence between divergent events on interacting lineages, *e.g.* cospeciation (Box 1.1). Finally, it is important to keep in mind that coevolution and codiversification are not caused by or evidence of one another and different mechanisms can result in one or a mix of both (Janzen, 1980; Segraves, 2010; Janz, 2011).

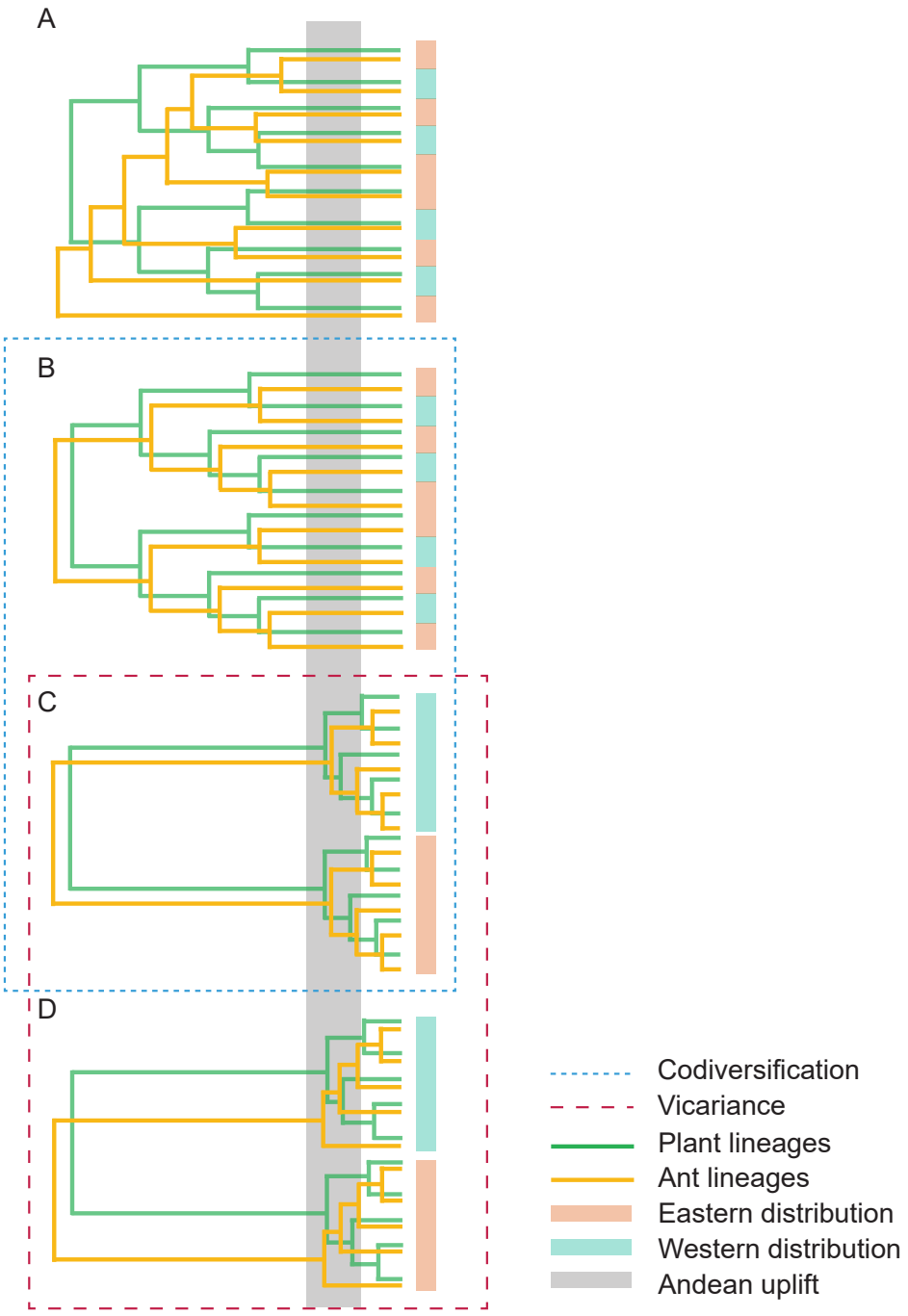


Figure 1.1: Expected patterns in the phylogenies of two mutualist organisms (in this case exemplified by ants, plants and the Andean uplift) arising from different evolutionary processes. Plant phylogenies are in green and ant phylogenies in orange. **A.** Topological and temporal incongruence between both phylogenies. The Andean uplift has had no effect on the evolution of either plants or ants. **B.** Codiversification (and possibly coevolution) resulting in topological and temporal congruence between ants and plants. Temporal incongruence with the Andean uplift. **C.** Topological and temporal congruence between plants, ants and the Andean uplift. In this model, distinguishing between codiversification, covicariance, and coevolution is difficult. **D.** Topological incongruence towards the tips of plant and ant phylogenies. Temporal congruence between the Andean uplift and both organisms. Here, covicariance and not codiversification is responsible for the phylogenetic patterns. Coevolution can occur, but in the absence of codiversification.

1.1.2 Codiversification in the presence of coevolution

Coevolution and codiversification can happen among lineages involved in any kind of association: mutualism, parasitism, predation or competition, and each lineage will adopt a term specific to its association; however, I will refer to one and the other as host and symbiont to minimize confusion (thus, symbiont here does not necessarily imply strict symbiosis like that which occurs between *Wolbachia* and ants). As coevolution implies a series of reciprocal adaptations arising between interacting lineages, it is easy to imagine that such host and symbiont tracking will also show signals of codiversification. It is possible for a host lineage to diversify in response to biotic or abiotic factors, and therefore inflict diverging forces upon the symbiont lineage. In vertically transmitted mutualisms, the symbiont is directly passed to the next generation of associates with almost no chance of exchange of symbionts between species or even populations, making them good candidates for codiversification (Segraves, 2010). If host populations are structured, the symbionts' phylogeny is expected to reflect the same structure, or the other way around if the symbiont determines the hosts' evolution (B and C in Figure 1.1). *Drosophila melanogaster* is commonly infected by a wMel *Wolbachia* strain that is inherited maternally and can alter the frequency of *D. melanogaster* mitochondrial haplotypes by affecting *D. melanogaster* reproductive success (Werren et al., 2008). A study including several *D. melanogaster* mitotypes distributed across the globe found that geographic structure of *D. melanogaster* mitotypes is correlated with different sub strains of wMel, evidence of coevolution and possible codiversification (acting at a population level) between *Wolbachia* and *D. melanogaster* (Ilinsky, 2013). Phylogenetic concordance and coevolution between *Wolbachia* and nematode hosts have been demonstrated; however, codiversification does not always occur between bacterial and

other arthropod hosts (Werren et al., 2008). Another example where coevolution and codiversification have been demonstrated is the mutualism between aphids of the genus *Brachycaudus* and the vertically transmitted *Buchnera aphidicola* which synthesizes amino acids missing from the phloem-based diet in aphids (Jousselin et al., 2009). Different strains of the bacteria are specific to different *Brachycaudus* and significance tests support the hypothesis of parallel speciation between the bacteria and the aphids.

Delayed codiversification, or phylogenetic tracking, can occur if coevolution facilitates lineage diversification due to the exploitation of new niches. However, some authors might not consider this case as an example of codiversification because the divergent events of the symbiont occur sometime after those of the host. Jumping plant-lice from the Psylloidea group are species specific to plants of the Genisteae tribe (Fabaceae) and it is known that both have undergone adaptive radiations in Europe and North Africa (Percy, 2003). Analyses of phylogenetic reconciliation and fossil calibrations found that plants from the tribe speciated at about 5-7 Mya likely because of the production of quinolizidine alkaloids as an herbivore deterrent. Only after the Psylloidea lice acquired mechanisms to overcome the toxin did they speciate and establish associations with host plants, around 2.9-3.4 Mya (Percy et al., 2004).

Coevolving populations with a wide distribution range can be geographically structured, which in addition to genetic drift, isolation by distance and differential gene flow between populations, can result in different magnitudes of coevolution and different adaptation regimes throughout the range of the association (Kiester et al., 1984; Thompson, 1994, 2005; Althoff et al., 2014). These variable selection schemes result in what Thompson calls hotspots and coldspots of coevolution, in the context of his proposed geographic mosaic theory of coevolution (GMTC, Thompson 2005). This model

can explain differences in patterns of coevolution and/or codiversification occurring in different populations of interacting species while accounting for ecological, historical and geographical local differences. Association-related traits undergo locally reciprocal phenotypic variations or local coadaptations. As species associations are geographically structured, trait differences (*i.e.* host/symbiont preference, habitat preference) evolve differently among populations of the same species. If population structure is sufficient and gene flow between interacting populations is restricted, local coevolutionary processes can lead to codiversification (Segraves, 2010).

1.1.3 Coevolution in the absence of codiversification

Coevolution does not always cause codiversification and does not require associations to occur between closely related sets of hosts and symbionts. The traits mediating the associations and upon which reciprocal selection acts can be the result of convergent evolution and not necessarily be restricted to a single taxon. Furthermore, patterns of coevolution in the absence of codiversification observed in phylogenies can result from the poor sampling of either one or both host and symbiont: extinct lineages or sampling error will likely produce phylogenetic incongruence (Page, 1993). In other cases where patterns of codiversification are detected, these are not always related to coevolution but to shared geographic backgrounds (as it is discussed later). Coevolution without codiversification can emerge from host switches, plausible mechanisms given phylogenetic niche conservatism. *Tegeticula* moths are species-specific pollinators of the *Yucca* plant (Aker and Udovic, 1981). They pollinate the flowers using modified mouth parts and feed on a portion of the fruits produced by the plant. Moreover, moths of the genus *Prodoxus* are commensals of the association and feed on the fruits

without pollinating the flowers. By comparing the phylogenies of *Tegeticula* and *Prodoxus* (a non-pollinator sister genus) moths to one of *Yucca* species, and reconstructing the distribution ranges for all three taxa, Althoff et al. (2012) found patterns of codiversification between *Yucca* and its specialist moth, as well as with the commensal moth. The reconstructed distribution range demonstrated that *Yucca* species within each lineage occurred in allopatry and that biogeographic factors were the main drivers of codiversification and not coevolution as previously thought (Althoff et al., 2012).

Despite the assumption that hosts and symbionts are highly specialized, host switching is a dominant factor influencing host-symbiont associations (Araujo et al., 2015). Selection is expected to be strong in horizontally transmitted associations as host and symbiont must recognize and select the best partner while avoiding detrimental outcomes. And because horizontal transmission requires the association to be established every time there is a new generation of hosts or symbionts, the chances of associating with a new partner species and expanding the host/symbiont range are higher compared to vertically transmitted associations. Ecological fitting and resource tracking are mechanisms by which partners can associate with species that, due to similarity, replace the sources of the ancestral partner (Araujo et al., 2015). Ants living inside plants are a good example of host switching. The habit of colonizing and depending on plants has occurred independently many times within the Formicidae family. Likewise, around 1139 species of plants from different genera and families, associate with different genera of ants in a mutualism that has evolved independently multiple times (Davidson, 1993; Davidson and McKey, 1993; Chomicki and Renner, 2015).

Detecting coevolution requires testing selection upon the traits involved in the association and a sufficient sampling of hosts and symbionts to be able to reveal signals of

it in phylogenetic studies. Reciprocal transplant experiments in which individuals of different partner taxa are swapped and the fitness of the new association measured are a good example of complement experiments for coevolutionary testing (Althoff et al., 2014).

1.1.4 Codiversification in the absence of coevolution

Cases of codiversification where there is no evidence of coevolution can be explained by geographic changes causing host and symbiont populations to split if both are distributed in the same region (Herre et al., 1999; Segraves, 2010; Hembry et al., 2014). Among the evidence favoring simultaneous vicariance due to shared geographic histories having a larger effect than coevolution is the temporal congruence between divergent and geographic events and the similar geographic structure patterns in taxa not related to the association. Even in the presence of low gene flow between populations that are isolated (by distance, for instance), codiversification can occur due to local selective forces acting differently between locations (Segraves, 2010).

Examples of speciation driven by geography are more abundant in individual taxa than in mutualisms as research on associations tends to test for coevolution or codiversification under the light of coevolution, not to prove or discard codiversification as a major driver of evolution in coevolving systems. Mutualisms are assumed to coevolve and attempts to address codiversification will, by definition, include coevolution. The question then becomes whether codiversification plays a greater role in mutualism diversification than coevolution does. To answer this the best systems to study would be those coevolving in areas where geographic changes have the potential for gene flow and

have occurred recently so their imprints are detectable and not confounded by genetic admixture.

1.1.5 The Andes cordillera

The Neotropics hold a high proportion of the worldś species diversity and are characterized by recent geographical changes and highly diversified taxa (Antonelli et al., 2010). Neotropical diversification likely resulted from a complex mix of biotic and abiotic mechanisms; however, Neogene tectonics and Quaternary climate cycles have played an important role in the origins of such diversity (Rull, 2011). The importance of the emergence of plant-insect associations (like pollination) and niche conservatism has been discussed as another potential driver of neotropical diversification (Antonelli and Sanmartín, 2011). According to Hoorn et al. (2010), most species of plants, mammals, birds, amphibians, insects and arachnids from the Amazon basin diversified during the Neogene, which dates from 23.3 Mya to 5.33 Mya (Hoorn et al., 2010). Fossil-calibrated molecular phylogenies showed that some of those species continued diversifying during the Quaternary, possibly driven by climatic and biotic changes (Rull et al., 2008; Rull, 2011). Many studies have proposed the Andean mountain chain as the driver of such speciation via allopatric speciation and ecological displacement, and by altering the hydrology and climate of the region (Brumfield and Edwards, 2007; Antonelli et al., 2009). But despite the implications that landscape changes might have on ecologically restricted taxa and their interspecific associations, only a few studies address the effect of geographic barriers in mutualisms and even fewer include more than one of the mutualists.

The Andean cordillera orogeny was not the result of a single event but of different episodes in time and space (Gregory-Wodzicki, 2000). The Andean Cordillera runs from southern Argentina and Chile to northern Venezuela and Colombia, where it splits into three mountain chains (Figure 1.2). The Andean northern block, encompassing Ecuador, Colombia and Venezuela, and specifically in Colombia, is where most of the tectonic activity has taken place since the late Mesozoic (Cediel et al., 2003). From the late Cretaceous to the Paleocene, the Western and Central cordilleras uplifted relatively slowly followed by a faster elevation rate from the Pliocene to the Holocene during which the processes that led to the formation of the Eastern Cordillera were most intense (Gregory-Wodzicki, 2000). Additionally, from the Miocene to the Pliocene most mountain elevations were approximately up to 40% of the modern values but experienced a rapid increase in elevation from 5 to 2 Mya (Gregory-Wodzicki, 2000). The Andes Cordillera provides a good scenario for testing the effects of geographic changes on the evolution of species and associations among them.

The Andes uplift caused rain regimes to change and precipitated the drainage of marine incursions on the Amazon Craton. A fluvial-marine system partially flooded the Amazon Craton and the rest of northwest South America during the early Miocene (23-10 Mya), due to marine incursions from the north and east coasts of South America (Hoorn, 1993; Hoorn et al., 1995) and increased rainfall in the east as the mountains became higher (Hoorn et al., 2010). These incursions were drained from west to east, presumably helped by the mountain uplift, and the mega-wetlands of the Lake Pebas system on the Amazon Craton dried as the Amazon river changed its course towards the end of the Late Miocene between (Hoorn, 1993; Hoorn et al., 2010; Shephard et al., 2010) (C in Figure 1.3). Other changes in prevailing wind direction, the source of

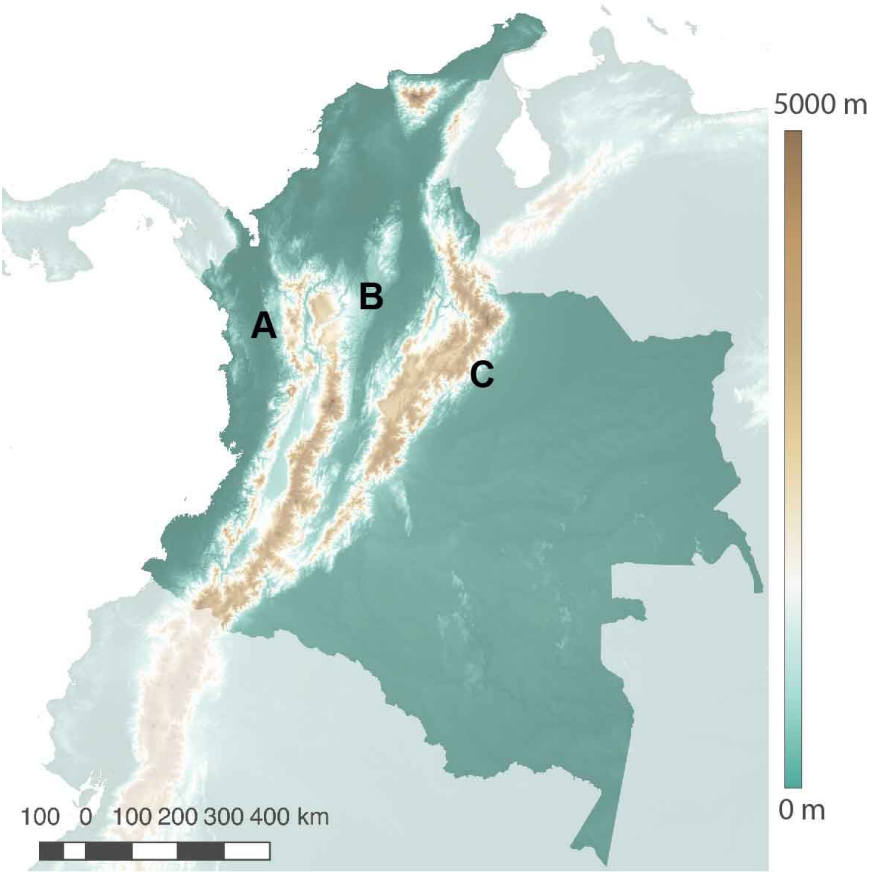


Figure 1.2: Topographic map of the Andes cordillera in Colombia. The Andes split in three: A. Western Cordillera, B. Central Cordillera, C. Eastern Cordillera. The Magdalena river flows through the Magdalena valley between the Eastern and Central Cordilleras, while the Cauca river flows between the Central and Western Cordilleras.

evaporated water for rainfall and surface temperatures are predicted to have happened because of the Andes uplift (Ehlers and Poulsen, 2009).

1.1.6 Myrmecophytism: ant and plant mutualisms

Ant-plant symbioses, or myrmecophytism, are commonly studied as examples of co-evolution and models to understand the evolutionary dynamics of beneficial interspecific associations (Davidson and McKey, 1993; Heil and McKey, 2003). Janzen (1966) described for the first time the associations between *Pseudomyrmex* ants and *Acacia* plants, demonstrating the benefits of the cooperation. Since then, obligate and

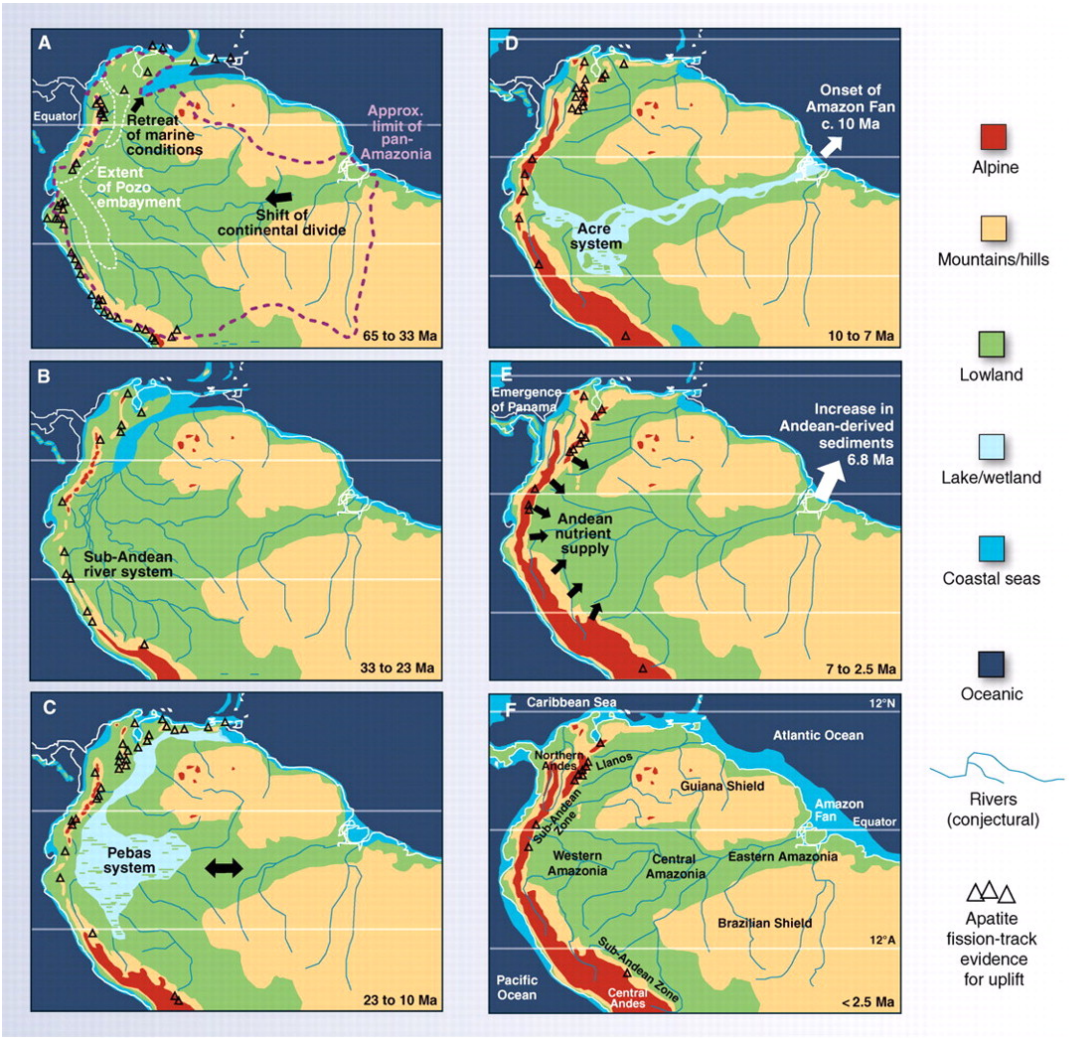


Figure 1.3: Paleogeographic maps of the Andean uplift, taken from (Hoorn et al., 2010). (A) Extension of Amazonia in the northern South America at the time the Andes started uplifting. (B) The Andes continued to rise. (C) Mountain building in the Central and Northern Andes at approximately 12 Mya and marine incursions forming the Pebas system in western Amazonia. (D) Northern Andean uplift which facilitated speciation. (E) The wetlands drained and the rainforests expanded. (F) South America's migration northwards during the Paleogene.

facultative myrmecophyte associations are documented as independently evolving between ants and plants, from ferns to angiosperms (Beattie, 1985; Davidson and McKey, 1993; Koptur et al., 1998; Rico-Gray and Oliveira, 2007). In the Brazilian rain forest alone Fonseca and Ganade (1996) reported around 337 myrmecophyte individuals per ha. High diversity regarding associated partners characterizes ant-plant mutualisms and determines cost-benefit trade-offs (Bronstein, 1998). As important components

of tropical communities, myrmecophytes play a key role in structuring food webs and maintaining diversity (Chenuil and McKey, 1996; Heil and McKey, 2003).

Ants (Formicidae) represent an estimate of 10-15% of the entire animal biomass in land (Beattie and Hughes, 2002), form eusocial colonies (Hölldobler and Wilson, 1990), and communicate among them and with their environment through chemical signals (Hölldobler and Wilson, 1990; Rico-Gray and Oliveira, 2007). Ants are highly diverse in tropical and subtropical areas, and their diversity only decreases with an increase in latitude, altitude and aridity (Beattie and Hughes, 2002). Moreover, their diversity is linked to several radiations, to multiple adaptations like eusociality and the presence of a metapleural gland that produces antibiotic fluids (Hölldobler and Engel-Siegel, 1984; Hölldobler and Wilson, 1990; Beattie and Hughes, 2002). Myrmecophytic interactions have been recorded mostly among angiosperms, but ant-fern interactions involving *Polypodium, Asplenium* and *Solanopteris* ferns associated with *Pheidole, Brachymyrmex, Leptothorax* and *Solenopsis* ants have also been recorded (Koptur et al., 1998; Rico-Gray and Oliveira, 2007; Fayle et al., 2011). From an estimated 15000 ant species (Bolton and others, 1994), around 110 species from 5 subfamilies are inhabitants of plants, a trait appearing several independent times across ant and plant phylogenies (Chomicki and Renner, 2015). The first evidence of an ant-plant interactions dates from the Eocene-Oligocene (35 Mya, million years ago) and it is the fossil of a *Populus crassa* leaf bearing Extra Floral Nectaries (EFN), used by modern angiosperms to attract ants and provide them with food (Pemberton, 1992). However, fossil evidence of ant-inhabited plant domatia is very scarce (Rico-Gray and Oliveira, 2007). In a phylogenetic study mapping the presence of domatia and EFN in plants, Chomicki and Renner (2015) estimate that domatia have appeared 158 independent times and have been lost about 43, and that the earliest domatia appeared 19 Mya in Australasia and

15 Mya in the Neotropics. They also predict the number of vascular plants bearing domatia to be 681.

Plant-associated ants occupy hollow cavities or specialized structures (called ant domatia, a, b, c, and e in Figure 1.4) and use food resources provided by the plant such as nectar produced in extrafloral nectaries, food bodies (*e.g.* Müllerian bodies produced by *Cecropia*) and glandular trichomes (f in Figure 1.4)(Davidson et al., 1989; Davidson and McKey, 1993; Rosumek et al., 2009). Often, ant colonies tend honey-dew coccids or pseudococcids (scale insects) on the host, and use them to feed on the plant's phloem and from it obtain sugar-rich secretions as an indirect way of getting food from the plant (Cabrera and Jaffé, 1994). Simultaneously, the ant colony acts as the host's biotic defense against herbivores and plant competitors (g in Figure 1.4) (Janzen, 1985; Alvarez et al., 2001; Bronstein et al., 2006). Aside from the benefits of hosting an army, some myrmecophytic plants can absorb waste deposited inside the domatia by the ants (Beattie, 1989; Treseder et al., 1995; Solano and Dejean, 2004). Radioactive evidence in *Tococa guianensis* has demonstrated absorption of organic waste from the ant colonies, which in turn feed on glandular trichomes containing lipids and sugars produced by the plant (Cabrera and Jaffé, 1994).

1.1.7 Establishment and dispersion of the myrmecophytism

Unlike many other symbiotic systems, ant-plant mutualisms are reassembled every generation via horizontal transmission, increasing the risk of colonists not re-locating hosts (Edwards et al., 2006). The process of host colonization starts after alates mate and queens identify their hosts, usually using volatile cues (Dáttilo et al., 2009; Torres and

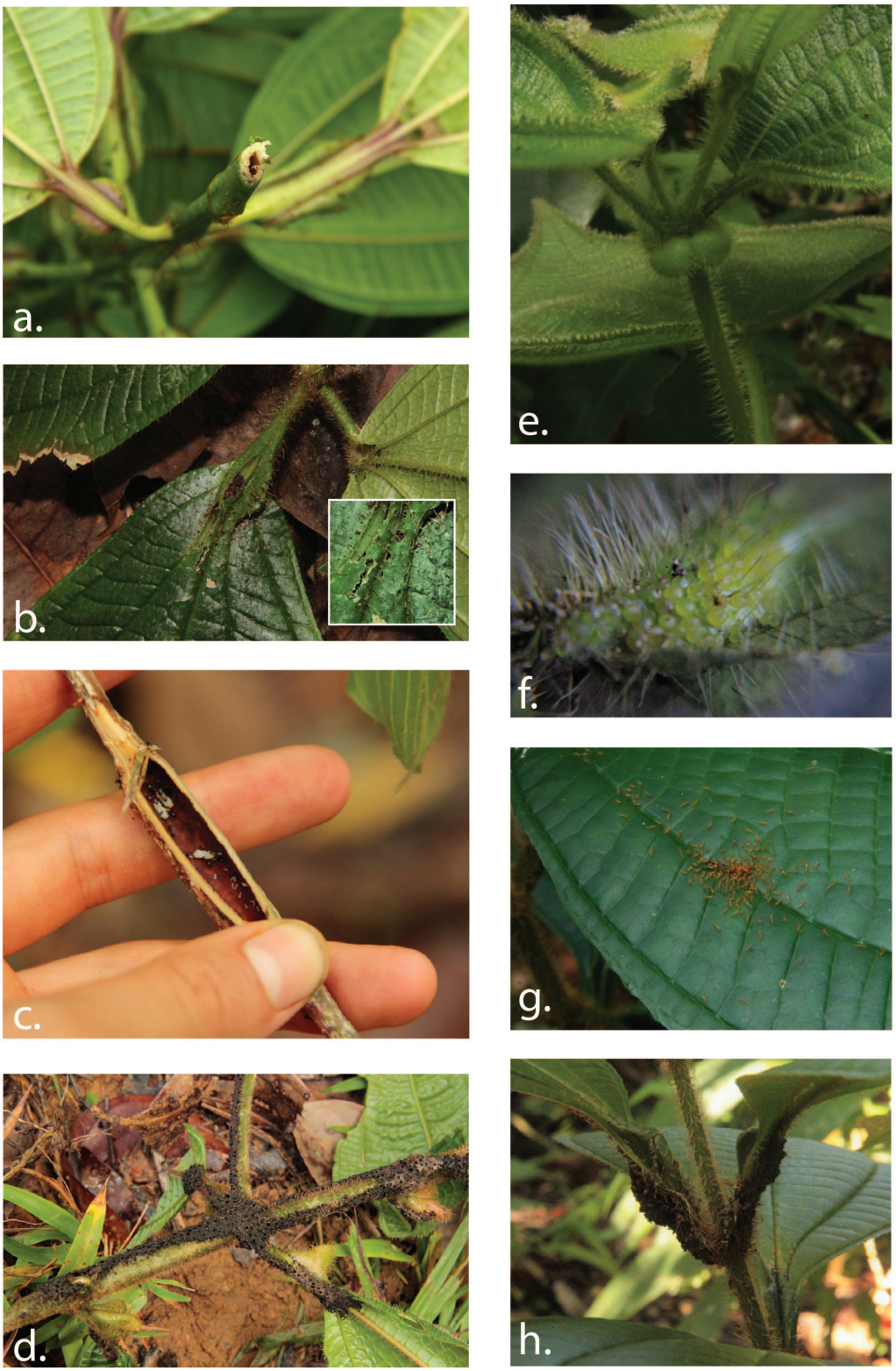


Figure 1.4: **a.** Rare caulinary domatium found in *T. guianensis* from Meta. Foliar domatia were occupied more densely. **b.** Domatia in *Maieta*, embedded on the leaf, where the ants inhabit. The caption shows entrances made by *Pheidole* ants. **c.** Opened caulinary domatium of *Duroia* inhabited by *Crematogaster*. **d.** Carton nest built by *Pheidole* ants. **e.** Domatia in *Clidemia*, mostly for mites. **f.** Trichomes in *Maieta guianensis*. **g.** *Pheidole* ants removing an insect carcass off the leaf blade. **h.** Carton nest built by *Pheidole* at the entrances to *M. guianensis* domatia.

Sanchez, 2017). Subsequent success at establishing a colony in a host and the composition of ant-plant communities will depend on host availability, distance to the host, and interspecific competition among queens Davidson et al. (1989); Heil and McKey (2003). For instance, *Azteca* and *Allomerus* ants have different strategies to compete for the colonization of *Cordia nodosa* in Western Amazon (Yu et al., 2001). *Azteca* queens can fly longer distances than *Allomerus* and colonize available hosts that are far away from the source colony. Although flying capacity is lower in *Allomerus*, these queens are more fecund and can displace *Azteca* from areas where host density is higher (Yu et al., 2001, 2004). Observations of colonized plants from very early stages of development and successful colonization events by multiple ant species are both evidence of strong intra and interspecific competition once the queens have found their host (Davidson et al., 1989). In myrmecophyte plants with modular or multiple independent domatia, every domatium can be occupied by different ant species. The competition is solved once the one colony is successfully established and its growing displaces its competitors (Davidson and Fisher, 1991).

Once the winning colony is established it remains in the same host during its lifetime. Similarly, myrmecophytic plants are continuously inhabited by ants during most of their lifetime (Webber et al., 2007; Pringle et al., 2014). During this process, fertile alate queens leave the colony and fly for long distances looking for a new host to colonize after mating takes place (Jürgens et al., 2006). But competition for unoccupied hosts and predation increases with distance and time. A trade-off between flight muscle size (that correlates with the ability to cover long distances) and host abundance and distribution are found among different plant-ants (Murrell et al., 2002; Bruna et al., 2011; Helms and Kaspari, 2015). On the other hand, unoccupied young plants might benefit from the rapid establishment of the defending ant colony to reduce herbivory

and competition. In contrast to ants, angiosperms can disperse long distances via seed dispersal by birds, although their survival depends mostly on the availability of defending colonies (Vasconcelos, 1991). Therefore, plant and ant distributions, density, and population structure are interdependent and restricted by the dispersal ability of their partners. Myrmecophyte associations are of interest as potential benefits of the association have been shown and the inter-dependency of one another seems clear in most cases.

Ant-plant mutualisms are ubiquitous associations with varying degrees of specificity potentially leading to different patterns of coevolution. For instance, generalist mutualisms have a lower expected potential for coevolution and codiversification than highly specific mutualisms. However, emerging patterns of codiversification might be erroneously attributed to coevolution, overlooking the potential role of abiotic factors in promoting codiversification. This is particularly likely when the mutualism is distributed over a large area that overlaps with potential barriers to gene flow for one or both associates, such as mountains or rivers. When looking at evolutionary patterns in the mutualism, evidence supporting codiversification due to shared geographic histories and in the absence of coevolution includes temporal overlap of diversification and geographic events (Page, 2003; Althoff et al., 2014), in addition to similar patterns of geographic structure in taxa not related to the association. Thus, to disentangle confounding factors causing convergent phylogenies between mutualists, it is essential to understand the role geographic history has upon the mutualism's evolution.

The Neotropical region is perhaps the most diverse region on Earth, partly due to recent geographic events isolating populations and increasing diversification and partly due to the establishment of interspecific associations. As organisms are not isolated

entities and associations of all kind occur among organisms, it is interesting to study to what extent either geographic or ecological history have influenced the establishment of these associations and how relevant they are as promoters of diversity. The Neotropics also hold the highest diversity of ant-plant mutualisms, with plant families Rubiaceae and Melastomataceae having 162 and 144 myrmecophyte species respectively (Chomicki and Renner, 2015). The *T. guianensis-Azteca* (Melastomataceae plants and Dolichoderinae ants) study system is interesting because of its wide distribution over an area that has dynamically changed recently. This permits testing of the effects of such changes over associations and, eventually, the relative contribution of these abiotic changes compared to biological causes of diversification. *T. guianensis* is distributed on both sides of the Andean cordillera while many other myrmecophyte species and genera have either an Amazonian or a Pacific distribution. Additional advantages of the system include a relatively robust fossil record (as used in Morley and Dick, 2003, Berger et al., 2016, and Moreau and Bell, 2011, for Miconieae and *Azteca*), and the fact that both genera include non-myrmecophyte species, which allows for comparative studies looking at rate differences between mutualist and non-mutualist lineages for example. Additionally, both *Azteca* and *Tococa* encompass highly diverse genera and belong to equally diverse families. Despite the system being promising for the study of mutualism evolution, the disadvantages of the *T. guianensis-Azteca* mutualism are related to species identification and phylogenetic resolution, typical of young and diverse taxa. To overcome these disadvantages, the general aim of my project is to uncover information about the system to develop it as a model to study the evolution of mutualisms and the evolution of ant-plant associations in general.

1.1.8 Ant genus *Azteca*

Azteca (subfamily Dolichoderinae) is a large Neotropical ant genus encompassing about 84 species exhibiting different nesting behaviors, including myrmecophytism, which has evolved independently multiple times within the genus (Longino, 1989). Extensive morphological differences among individuals in different localities suggest species diversification throughout the South American tropics (Longino, 1991c,b). *Azteca* is commonly associated with *Tococa* (Melastomataceae), but also nests in *Cordia* (Boraginaceae), *Tachigali* (Fabaceae), *Cecropia* (Urticaceae) and occasionally *Triplaris* (Polygonaceae) (Longino1991a). Additionally, a strong habitat-specialization rather than host-specialization was previously reported for *Azteca*, likely related to resource availability and herbivore defense issues (Longino, 1991c; Yu and Davidson, 1997). Moreover, experiments demonstrate the good dispersal capability of *Azteca* in comparison with other ant genera in terms of wing muscle size, but estimate the average distance coverage is between 400-500 m (Bruna et al., 2011). Although some ant species can disperse extensive distances over flat terrain, large altitude differentials limit dispersion across mountains.

1.1.9 *Tococa guianensis*

The Melastomataceae family includes 4079 accepted species distributed among eight tribes, from which 11 genera are myrmecophytes, including *Tococa* (Renner, 1993; Michelangeli, 2010a). Among Melastomataceae tribes, thirteen genera in the Tropical Americas are myrmecophytes (Michelangeli, 2010a) and other two are facultatively associated with ants (Clausing, 1998). Hollow structures for hosting associates (ants and occasionally mites, as is the case for *Miconia* and *Blakea*) are present in the shape

of petiolar sacs (*e.g. Tococa*, *Maieta*) or hollow stems (*Miconia*) (Michelangeli, 2010a). These myrmecophytes establish associations with a wide range of ant genera. For instance, within tribe Miconieae, *Miconia guianensis* associates with *Pheidole* and *Crematogaster* ants (Vasconcelos, 1991; Morawetz et al., 1992; Lapola et al., 2003), and *Tococa guianensis* with *Azteca* and *Pheidole* (Alvarez et al., 2001; Bizerril and Vieira, 2002; Michelangeli, 2003). Others like *Tococa macrosperma* associate with *Allomerus* and *Crematogaster* ants (Michelangeli, 2003). Within tribe Blakeeae, *Blakea* and *Topobea* are presumed to host ants and mites into hollow internodes, layered stipules, and leaves (Renner, 1989; Penneys and Judd, 2011), but the identity of their associates is not reported. More facultative associations are observed in *Pachycentria constricta* and *P. glauca*. Both species can grow within ant nests and produce seeds that are attractive to (and dispersed by) them (Clausing, 1998, 2000). A similar interaction between the species *Medinilla speciosa* and *Dolichoderus* ants is mediated by the production of pearl bodies to feed the ants and hollow root swellings for hosting them (Clausing, 1998, 2000; Clausing and Renner, 2001).

Ants do not trigger the production of the leaf sacs. These are instead a preadaptation of the plant in which a specialized tissue grows at the insertion of the blade into the petiole (Bitallion, 1982; Alvarez et al., 2001). *Tococa guianensis* (Figures 1.6 and 1.7, Melastomataceae) is one of the most widely distributed species of *Tococa*, commonly found from Central to South America (Michelangeli, 2005). Ants nest in domatia placed in the petiole, the blade of the leaf, or the hollow stems (Michelangeli, 2010a). Particularly associated with *Tococa* are the ant genera *Azteca*, *Crematogaster*, *Allomerus*, *Brachymyrmex*, *Paratrechina*, *Solenopsis*, *Wasmannia* and *Pheidole*, though *Pheidole* and *Azteca* constitute more strict partners than the others (Cabrera and Jaffé, 1994;

Michelangeli, 2005, 2010a). *Myrmelachista* ant species prune the vegetation surrounding their host, resulting in monospecific plots in the forest known as devil's gardens (Renner and Ricklefs, 1998). In an experiment of ant exclusion, Michelangeli (2003) found that inhabiting ants protect three Amazonian *Tococa* species from herbivory (Figure 1.5). Scale insects, and possibly nitrogen-fixing bacteria, are also involved in the mutualism. *Tococa* produces glandular trichomes, which produce lipid- and sugar-rich rewards for the colony (Alvarez et al., 2001), and the plants can use the ant waste products as a source of nutrients, particularly nitrogen (Solano and Dejean, 2004).

Tococa shrubs normally grow in humid places near water bodies and most of the species are distributed in the lowlands on either side of the Andes, rarely above 1,200 m.a.s.l. This restriction allows us to predict in which localities it might be found (Goldenberg et al., 2008; Michelangeli et al., 2008). At the first developmental stage of *T. guianensis* the only two leaves of the seedling lack domatia, which appears first in only one leaf on the next pair and in both leaves in the following pair and thereafter (Alvarez et al., 2001).

1.1.10 Hypotheses

Tococa guianensis and *Azteca* distributions across the Andes overlap and they share the geographic history of the region. Moreover, associations are not genus-specific as both can associate with other plant or ant genera. Thus, the hypothesis is that the uplift of the Andean mountains had a significant effect on the evolution of both lineages producing vicariance events congruent with the uplift of the Andes, as oppose to codiversification events occurring before or after the uplift (*e.g.* via post-uplift dispersal). To assess the hypothesis, this project aims to look for evidence of incipient divergence

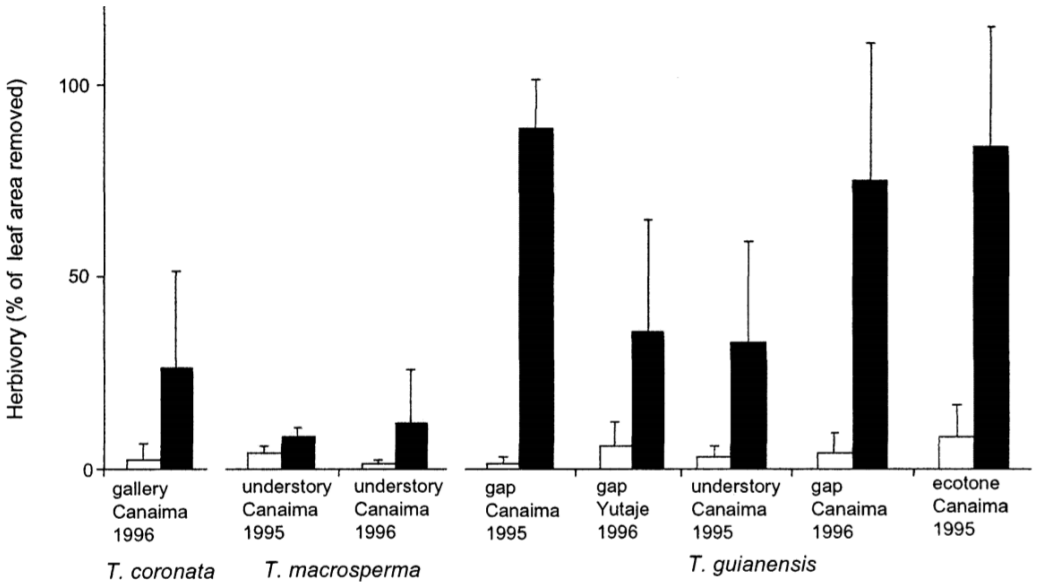


Figure 1.5: Results of the ant-exclusion experiment for eight populations of *Tococa*, including *T. guianensis*. Empty bars represent control populations and solid bars represent experimental populations from which ant colonies were removed. Figure taken from Michelangeli (2003)

Figure 1.6: *Tococa guianensis* inflorescence. Leaf domatium are visible in the background

Figure 1.7: *Tococa guianensis* domatium with dead *Azteca* ants inside.

between populations in the shape of population structure and absence of gene flow and to test whether geography rather than codiversification is the mechanism behind the patterns of evolution observed in *T. guianensis* and *Azteca*. It additionally aims to resolve the taxonomic uncertainty of each taxon to explore the phylogeographic patterns of the ant-plant mutualism. To do this, I need to answer the following questions: *1.* Which *Azteca* ants are associated with *T. guianensis*, and are they the same throughout the distribution of *T. guianensis*? *2.* Is the Andean cordillera acting as a barrier to gene flow, separating populations on one side from the other and therefore potentially contributing to population divergence? *3.* Did the Andean cordillera act as a barrier to gene flow in the past, producing vicariance patterns in both? *4.* Does evidence suggest the same ant's evolutionary patterns hold for the ant's vertically transmitted symbionts? Simultaneously, the project aims to produce genomic data from *T. guianensis* of sufficient quality to facilitate further analyses, including the identification of variable sites that can contribute to the resolution of the Miconieae tribe.

1.1.11 Alternative hypotheses

One alternative hypothesis is that the uplift of the Andes had effects on the ant lineages but not in the plant lineages. Based on dispersal limitations ants encounter and the short distances they can cover compared to those of plants, I expect to observe highly geographically structured ant populations, having more than one differentiated population on each collecting site, while having a more homogeneous, perhaps single, plant population across the Andes. Moreover, I expect from ant population structure to be determined by the Andes Cordillera as an indication of the mountains acting as a barrier to gene flow between ant populations. As *Tococa* plants are dispersed mostly by birds, I expect geographically close populations to be genetically more similar than populations isolated by a large distance, regardless of the mountain range. If ant and plant populations differ in their tree topologies and the direction and magnitude of gene flow, then this constitutes evidence against codivergence and suggests that vicariance and shared geographic history is more relevant. In addition, if the Andean Cordillera is a stronger barrier to gene flow in ants compared to plants, then they will not have a shared geographic history.

Other alternative hypotheses include finding patterns of codiversification that do not reflect the geographic history of the area. In this case, coevolution could be a potential cause. A key innovation that evolved in an association because of coevolution can facilitate the geographic expansion of such association and the isolation of extreme populations (Thompson, 2005) (*e.g.* B in Figure 1.1). As a result, isolation by distance, genetic drift and new local adaptations will promote diversification. Therefore, another scenario of coevolution as the most important promoter of diversification can be to assume that the mountain ranges are not absolute barriers to gene flow, but that

diversification is an indirect result of coevolution. In this case, populations will be genetically similar to nearby populations regarding the presence of geographical barriers and will be less similar to distant populations. In other terms, more gene flow can be expected between adjacent populations than between distant populations (isolation by distance). If this is true, it can be evidence that the mutualism as a more relevant driver of diversification than geography, but further analyses will be required. Another scenario can be codiversification reflecting the geographic history of the area, in which case time-calibrated phylogenies will be necessary to address absolute time congruence between ants' and plants' population divergence and geographic events (C in Figure 1.1).

Expected patterns between ant and ant-symbionts will be of strict codiversification since the symbionts are vertically transmitted and only rare cases of horizontal transmission are possible (although these have been reported between *Wolbachia* and *Drosophila*). Evidence favoring this scenario will be to find different symbionts in each population and similar patterns of genetic divergence among them. An alternative scenario will be to find no structure in the symbionts' population related to the host or to geography but similar symbiont compositions across all ant populations, which can be an indicator of selection. From the plant point of view, the expectation is to produce genomic data that will further help developing primers for deeper and standardized sequencing.

Similar studies on the phylogeography of ant-plant mutualisms have been done; however, few studies address both taxa simultaneously and even fewer simultaneously collect individuals interacting in different, sparsely distributed geographic populations. From a literature review performed in PubMed, from 415 papers including the terms "phylogeography" and "plant* and ant*, or myrmecoph*", only 24 studies had an evolutionary

focus, included at least one phylogeny of one of the taxa and presented directly or indirectly geographic information (A includes see a summary of the 24 studies and the methods for the literature review). This project is one of the few addressing evolutionary hypotheses in a phylogeographic context simultaneously using ant and plant populations currently associated.

1.1.12 Project structure

The chapters in this thesis are a progression of work aiming at the identification of the subjects of a mutualism, the generation of data to determine relationships among subjects and to study the subject's evolution within a geographical framework. Chapter 2 uses a DNA barcoding approach to identify and delimit operational taxonomic units of *Azteca* ants collected from *T. guianensis* in the locations surrounding the northern Andes Cordilleras. In addition, I present fossil calibrated mtDNA and nDNA phylogenies. In contrast to *Tococa*, *Azteca*'s phylogenetic position is clear, but like the plant, morphological classification is difficult. The chapter establishes the baseline to select appropriate and comparable samples from each population to use in further analyses. Chapter 4 addresses the pitfalls of *Tococa* identification and species status and describes the assembly process of whole genomes of *T. guianensis*. These are the first whole genome sequences obtained from the Melastomataceae family and will complement currently available information. Chapter 3 describes the split assembly of the *Azteca* whole genome assemblies, evaluates divergence between ant populations distributed on both sides of the Cordillera, and tests the hypothesis that the mountains act as a barrier to gene flow and consequently are a possible driver of diversification. Results from this section will be compared with equivalent analyses on the host plants

to test for congruent evolutionary histories, signals of coevolution, codiversification and absolute time concurrence. This chapter also briefly evaluates *Azteca*'s *Wolbachia* endosymbionts.

A major limitation of this project has involved the difficulty of predicting the quality of data yielded from plant samples; however, obtaining reliable whole genome sequences is a good step forward. Regarding the ants, the limitations concern sampling and delimitation of clusters. Targeted sampling is impossible since the ant identity can be determined only after collecting the plant, and once collected, morphological classification based on worker ant's is not always accurate. In terms of hypotheses testing, it is not possible to make statements about signals of coevolution without a fully resolved phylogeny of both parties and without reciprocal experiments to test for selection in traits related to the association. What is possible, however, is to make inferences in the timing of the association's emergence based on what is known about the colonization of both genera into the area. The aim here was to first generate data to solve basic questions about the system and then explore effects of abiotic factors on the system's evolution. Once more data become available, hypotheses regarding the evolution of the mutualism can be tested.

CHAPTER

2

BIOGEOGRAPHY, DNA BARCODING AND DELIMITATION OF MOLECULAR OPERATIONAL TAXONOMIC UNITS IN *AZTECA* (FORMICIDAE: DOLICHODERINAE)

2.1 Introduction

This chapter sets out to identify the diversity of *Azteca* ant species associated with *Tococa* across Colombia, and specifically to identify dominant ant taxa that will be found using molecular analyses. It also places the timescale of divergence of *Azteca* ant species associated with *Tococa* in the context of the uplift history of the North Andean cordilleras. In the Introduction, I first lay out the challenges associated with ant (and particularly *Azteca*) identification in species-rich neotropical habitats. I then outline the value (and some limitations) of DNA barcoding approaches to the separation and definition of taxonomic units in organisms generally, and in *Azteca* ants more specifically. I provide a brief outline of the processes and timescale associated with the topography of Colombia, as background to my spatial sampling scheme. Finally, I outline my specific aims in more detail.

2.1.1 Ant diversity

Despite their individual size, ants represent an important portion of the world's biomass and are present in most terrestrial ecosystems. In addition, ants are highly diverse and exhibit rapid responses to environmental changes (Longino et al., 2002). As eusocial insects, ant adults are either reproductive queens, fertile males (drones) or non-reproductive female workers, and more than two generations can overlap in time in a single colony (Hölldobler and Wilson, 1990). The range of niches exploited and their varied habitats (*e.g.* arboreal or subterranean, with single or multiple queens per colony) reflect their remarkable diversity (Hölldobler and Wilson, 1990). Close to 16,000 valid species from 23 subfamilies and 512 genera are reported in AntWeb v.6.49 (`https://www.antweb.org`, accessed 17 April 2017). The origins of ant diversity are

often attributed to the rise of eusociality and the expansion of feeding strategies (Wilson and Hölldobler, 2005), in addition to the emergence of angiosperm-dominated forests (Moreau et al., 2006; Moreau and Bell, 2011, 2013).

More ant genera and species are found in the Neotropics than in any other region, including most endemic species restricted to tropical areas (Fisher, 2010). While some neotropical areas have a long history of traditional taxonomy (*e.g.* Janzen and Hallwachs, 2011), most have little local expertise in modern taxonomic research or ecological sampling. Moreover, cryptic diversity, higher taxonomic efforts on birds and mammals compared to insects, and just general lack of sampling (Stork, 2018). As a result, taxonomic keys do not exist for many ant taxa, and many species in biodiverse areas probably remain to be discovered and described. Stork (2018) estimates that around 80% of insect species, including ants, remain undiscovered. Given all of this, the scale of the taxonomic problem in ants is enormous.

2.1.2 Ant identification challenges

Morphologically cryptic species represent a major challenge for the taxonomic assessment of diverse tropical faunas (Hebert et al., 2004a). These are taxa that are hard or impossible even for professional taxonomists to separate reliably based on morphological criteria, but which can be identified as distinct species-level units using (most often) molecular approaches. Thus, cryptic species are common in taxa for which characters are difficult to categorize, see, or are simply not morphologically different (Bickford et al., 2007). For instance, cryptic species can occur in lineages exhibiting phenotypic plasticity, as such plasticity can derive in ecological speciation and genetic differentiation that does not produce morphological changes. In this case, diagnostic characters

vary in a continuum difficult to categorize by taxonomists. Similarly, it is possible that differences among species involve nonvisual mating signals that taxonomists cannot easily use (Bickford et al., 2007). In a study comparing *Formica japonica* ant colonies, authors found few morphological characters distinguishing cryptic species and that they differed in the types of cuticular ant-recognition hydrocarbons (Akino et al., 2018). Lastly, it is possible that divergence between cryptic species occurred recently and morphological differences have not accumulated, or that divergence occurred in a scenario where selection towards morphological characters is strong (Karsten et al., 2008; Bickford et al., 2007).

Incorrect identifications and failure to identify cryptic species are common when dealing with groups with large range distributions and whose morphological characters are uninformative (Hebert et al., 2003b,a; Seifert, 2009). Complex population differentiation and speciation processes, sometimes driven by hybridization or endosymbionts, mean that cryptic biodiversity is particularly common in ants (Paknia et al., 2015). The morphological structure of ant workers is conserved and simplified (meaning that there are relatively few diagnostic traits compared to, for example, the wing patterns of butterflies), and variation often involves continuous (rather than discrete) traits that increase the challenge for morphology-based taxonomic identification (Ross et al., 2009; Blaimer, 2012), particularly of closely-related sister species (Blaxter, 2004). In addition to the low reliability of species identification using workers, easier to classify queens are more difficult to collect as they are only seasonally present outside the nest (Longino, 2007; Cardoso et al., 2012). In the case of *Azteca*, the genus is a complex in which the range of character variation is higher than observed in a single species, such variation is partially discontinuous and suggests the existence of several species, geographic continuity is not well established and there might be not enough material to define species

(Longino, 1996). Moreover, workers and males exhibit continuous size polymorphism within and among colonies and species (Longino, 1996).

Given increasing recognition of ant's ecological importance, more effort has been invested in studies of ant diversity and species turnover than in producing formal species descriptions and the development of taxonomic tools (Bolton, 2003; Meier et al., 2006). Seberg (2004) predicted that at the current rate it would take about 940 years to finish describing all species known at that time. Hence, a growing number of studies are using DNA barcode sequences to sort and identify ant specimens (Smith et al., 2005; Smith and Fisher, 2009; Ngéndo et al., 2013; Smith et al., 2014; Paknia et al., 2015), and many other taxa (Hebert et al., 2003b; Ratnasingham and Hebert, 2013).

2.1.3 DNA barcoding

DNA barcoding (sequencing of one or more specific gene regions using highly-conserved DNA PCR primers) has been proposed as a rapid and cost-effective solution for specimen separation and identification (Hebert et al., 2003b; Kress and Erickson, 2008). Comparing sequences of unidentified specimens with reference sequences for morphologically identified voucher individuals facilitates rapid identification (Hebert et al., 2003b, 2004a; Blaxter, 2004; Edwards, 2009). The sequence region(s) used in DNA barcoding varies among major groups of organisms, depending on which markers, with highly conserved primers, have been found to be informative at the species level. For instance, the nuclear ITS2 locus is widely used in plants and animals (Chen et al., 2010; Yao et al., 2010), the Folmer region of the mitochondrial COI is the most widely used barcode locus in animals (Hebert et al., 2003b; Waugh, 2007), chloroplast genes matK, ycf1 and rbcL are widely used in plants (Newmaster et al., 2006; CBOL et al.,

2009; Dong et al., 2015), and nuclear 16S is the commonest barcode locus in Bacteria (Hajibabaei et al., 2007; Rosselli et al., 2016). Chloroplast and mitochondrial genes can have the advantage of a lower effective population size compared to nuclear markers (1/4 that of a nuclear marker, through being haploid and maternally inherited), and so generally are more sensitive to, and so more resolving of, the population bottleneck events that can accompany speciation. As DNA barcoding is applied more widely, increasing numbers of cases are revealed in which the use of multiple barcode loci is necessary to provide adequate resolution of identification (particularly in plants: *e.g.* Fazekas et al. (2008)) or to avoid misleading identification (for example, CO1 and a nuclear locus in some insects Nicholls et al. (2012), and see below). Many studies have nevertheless demonstrated the utility of DNA barcoding in inventories of hyper diverse organisms, identification of species complexes, the discovery of cryptic species, and rapid inventories (Smith et al., 2005; Tänzler et al., 2012; Cornils and Held, 2014; Hamilton et al., 2014). The method is based on the empirical observation (and the resulting assumption) that intraspecific divergence for a homologous gene is lower than interspecific divergence, such that if a query sequence is compared to a database (such as the BOLD Barcoding of Life database, or NCBI Genbank) it is possible to determine its taxon membership by quantifying the divergence between the query and the reference. Under ideal circumstances, intraspecific and interspecific distances have non-overlapping distributions, creating what Hebert et al. (2003b) called the Barcoding Gap. This can be defined as a ratio between the two distances (an average interspecific distance at least ten times larger than the average intraspecific distance) or in terms of threshold percentage sequence differences expected within or between species (see below) (Hebert et al., 2003a).

Collins and Cruickshank (2013) identify three benefits of DNA barcoding: specimen identification, species discovery and species delimitation. Specimen identification is the assignment of a taxonomic name to an unknown specimen by comparing it against reference sequences (Collins and Cruickshank, 2013). Assuming a reliable reference database, barcoding has advantages over adult-based morphological taxonomy in that it can be used when only fragments of a specimen are available for identification, when a life stage that has no taxonomic resource is all that is available (such as an immature stage) or to investigate whether a product is derived from an endangered species (Mitchell, 2015; Mendoza et al., 2016). Species discovery is akin to sorting specimens into species-like units (usually termed molecular operational taxonomic units, or MOTUs) by means of the genetic distances among specimens and using a single locus. It is widely accepted that without reference to sequenced voucher specimens, MOTUs are not equivalent to species, and that barcoding gaps may reflect the behavior of the few barcode loci and not the complete species history (Brower et al., 1996; Schindel and Miller, 2005; Brower, 2006; Rubinoff et al., 2006; Fujita et al., 2012). Finally, species delimitation refers to choosing a threshold to define boundaries of taxon status (species, population, etc.), using a multilocus and integrative approach that includes two or more loci from different sources (*i.e.* plastid, mitochondrial, and nuclear DNA), in addition to other molecular tools. These approaches reduce the bias associated with separating taxa using data for a single locus (see below) (Collins and Cruickshank, 2013).

2.1.4 Clustering and identification of specimens

Multiple methods have been developed to delimit taxa present within a set of sequences. Some use the barcoding gap as a single value threshold below which divergence is considered intraspecific and above which it is considered interspecific, *i.e.* a limit between the two taxonomic levels (Hebert et al., 2003a). Methods of species identification and delimitation rely on the discovery of this gap or threshold and can be classified depending on whether they require *a priori* defined groups or not. Early methods, such as **TaxonDNA** (Meier et al., 2006), were based on finding the barcode gap by computing all intra- and interspecific distances and finding the values at which the two distributions do not overlap, which requires having previous knowledge about the membership of the reference groups. Then, the unknown sample is assigned to the group to which its distance is lower than the barcode gap. This method is useful when the objective is to place a few query sequences within well-established groups:–for example, to identify products derived from endangered species or when specimens are damaged and identification cannot otherwise be achieved. Other methods can infer the barcoding gap or distance threshold from the data by doing all- against- all pairwise comparisons and generating a distribution of the genetic distances often with an additional validation step (**jMOTU** and **ABGD**). The advantage of these methods is that they are applicable to data from undescribed or unreferenced taxa. More complex methods incorporate likelihood and Bayesian frameworks or multispecies coalescent models (**BPP**, **GMYC**, **PTP**, **DISSECT**). Their advantage is the incorporation of uncertainty and a higher robustness to sampling bias and the number of loci used; however, some require an accurate 'guide tree' that can be difficult to obtain (**GMYC**, **PTP**) (Box B.1, in Appendix B).

2.1.5 The limitations of single locus DNA barcoding

Using barcodes for specimen identification and species delimitation has caveats that require attention, related to one or more of the limitations of reference databases, the number and mode of inheritance of barcode loci, and the validity of taxon definition thresholds when extrapolated to other data. First, many taxonomic groups have reference sequences matched to voucher specimens for only a tiny minority of species, particularly in biodiverse regions. In the absence of reference sequences, specimens can be assigned to MOTUs that can provide a basis for formulating species hypotheses for further confirmation. Second, using a single barcode locus can result in misidentification due to processes preventing sequences for any two species from forming discrete, non-overlapping, groups (Shaw, 2002; Rubinoff and Holland, 2005).

Barcoding works best when sequence sets for sister species are reciprocally monophyletic, though it can work when one taxon is paraphyletic with respect to a second if sequences sets are discrete. Problems in animal barcoding using the mitochondrial CO1 locus include *(i)* the existence of degenerate nuclear copies of mitochondrial sequences (NUMTs) that must be identified and excluded (Bensasson et al., 2001; Ballard and Whitlock, 2004), *(ii)* sharing of sequences between taxa due to incomplete lineage sorting and retention of shared ancestral polymorphism (Funk and Omland, 2003; Ballard and Whitlock, 2004; Rubinoff and Holland, 2005), *(iii)* transfer of mitochondrial genes between species through introgression during hybridization (Rubinoff and Holland, 2005), and *(iv)* selective sweeps on mitochondrial sequence variation imposed by maternally inherited symbionts. For instance, *Wolbachia* infections cause cytoplasmic incompatibility between infected and non-infected insects ultimately favoring reproduction with infected females over uninfected ones, thus selecting one mitochondrial

haplotype over another (Dean et al., 2003; Smith et al., 2012). These processes all result in mitochondrial DNA barcodes being non-monophyletic with respect to biological species –a pattern that is commonly observed in nature (Funk and Omland, 2003; Meyer and Paulay, 2005; Jansen et al., 2009; Nicholls et al., 2012; Paknia et al., 2015). Using more markers and contrasting mitochondrial against nuclear phylogenies can reveal incongruities caused by these phenomena and provides information of the lineages rather than the gene alone, reducing the chances of misidentification (*e.g.* Nicholls et al., 2012). High genetic structuring within a species (for example, through restriction of populations to different Pleistocene glacial refugia) can also result in intraspecific population structures that can mimic those between species (Lohse, 2009). Despite widespread use of a single locus in animal DNA barcoding, these limitations highlight the value of extensive geographic sampling and a multilocus approach that can cope with lineage sorting, past migration events and incongruent gene genealogies (Edwards, 2009). And even if methods are proven to be robust when using a single gene, if that gene is in the mitochondria or chloroplast, it is advisable to have at least one nuclear marker (Elias et al., 2007). Third, mutation rates, population sizes, and divergence times vary among species, which limits the use of universal thresholds and might introduce errors in the interpretation of these values (Meyer and Paulay, 2005; Yang and Rannala, 2016). When analyzing results, including external information about the specimens' distribution, ecology and natural history helps to assess the biological feasibility of the clusters.

2.1.6 DNA barcoding of *Azteca* ants on *Tococa*

The genus *Azteca* comprises 84 described species and 28 subspecies (AntWeb, `https://www.antweb.org/`, April 26th, 2017), which exhibit a variety of nesting habits that range from carton nests and the use of dead plant material to the use of live stems or other plant organs (Emery, 1893, 1913; Forel and Ogden, 1928; Longino, 1986, 1991c). Despite being one of the most diverse genera within the Dolichoderinae subtribe (Solvestre et al., 2003), taxonomic work on the genus remains very incomplete. *Azteca* was created as a temporary genus to place the *Liometopum* type, morphologically similar to both *Liometopum* and *Iridomyrmex* but differing in the gizzard and the number of casts (Forel, 1878). Placed in Dolichoderidae by Ashmead (1905), *Azteca* was later synonymized with *Liometopum xanthochroum* by Dalla Torre (1894) who included only four species on his ant catalogue. Wheeler (1912) synonymized *Azteca* and *Tapinoma* based on a *A. instabilis* type but was later corrected by Emery (1913) who kept *Azteca* as a genus and included it in the tribe Tapinomini (subfamily Dolichoderinae). Later, Shattuck (1992) places 130 *Azteca* species back to Dolichoderinae and out from the tribe Tapinomini, and highlights the poorly understood species boundaries within *Azteca*, identifying as possible causes the polymorphism and geographic variation exhibit by workers and many species. Shattuck's review (1992) is the last most comprehensive review of the genus and since then the number of accepted species has varied. Further taxonomic efforts have focused on clades associated with relatively well-studied neotropical ant-plant associations with *Cecropia* and *Cordia* plants, and the geographic scope of this work is limited mainly to Central America, Panama and Brazil (Longino, 1991c,a, 1996, 2007; Guerrero et al., 2010).

Phylogenetic work previously done in Dolichoderinae ants in addition to *Azteca* includes

phylogenies of the subfamily using one to three *Azteca* species and COI, 18S, 28S, wg markers (Chiotis et al., 2000; Ward et al., 2010). Similarly, Chiotis et al. (2000) used COI, COII, and Cytb from *A. longiceps* to place *Azteca* within Dolichoderinae and explore secondary structures of those markers. An additional cladogram was estimated for 21 genera of Dolichoderinae ants using 104 morphological characters and one undetermined *Azteca* species (Shattuck, 1995). For *Azteca*, Ayala et al. (1996) published a phylogeny of COI sequences from eight species associated to *Cecropia* and *Cordia*. At the population level, Debout et al. (2007) developed 12 microsatellites and measured population heterozygosity. Those microsatellites were further used to explore *A. instabilis* populations in Chiapas-Mexico (Remfert, 2012). *A. instabilis* does not form a strict mutualism with plants; however, it inhabits tree hollow trunks with fissures on them (Longino, 2007). Barriga et al. (2015) looked at ant-plant communities in Peru, Ecuador and Costa Rica, identifying *Azteca* queens morphologically and barcoding *Azteca* workers using the COI marker when queens were unavailable (Barriga et al., 2015). They defined seven *Azteca* MOTUs and identified seven referenced species. Finally, Pringle et al. (2012) used a concatenated matrix of one mitochondrial and four nuclear markers to reconstruct the phylogeny of nine morphologically identified *Azteca* species associated with *Cordia alliodora* (Boraginaceae) distributed from Costa Rica to Colombia, although only one collection from Colombia is available. From all their samples they identified nine *Azteca* species and five *A. pittieri* populations.

From the current 84 accepted species, ten and five recognized species are represented by sequences available in NCBI and BOLD respectively, mostly ITS2 sequences. From these, only one accession is from Colombia (both databases consulted on September 05, 2017). But despite the low representation of the diversity of *Azteca* in databases, molecular approaches can potentially provide enough resolution to distinguish between

Azteca species, as proven by Barriga et al. (2015) and Pringle et al. (2012). Moreover, DNA barcodes are more advantageous for discriminating between species when taxonomic revisions of local species are not available, especially when morphology varies geographically.

2.1.7 The geographic history of the Northern Andes

The placement of the timescale of *Tococa*-ant diversification into a regional topographic context requires a temporal hypothesis for the orogeny of the Northern Andes. The Andes Cordillera extends up the western side of South America from Chile to Colombia, where it splits into three geographically separated mountain ranges named the Western, Central and Eastern cordilleras (Figure 1.2 in Chapter 1). Caused by the subduction of the Nazca plate underneath the South American plate, the uplift of the Western and Central cordilleras started slowly in the Paleocene approximately 63 Mya (Million years ago) (A in Figure 1.3). Activity accelerated around 23 Mya and most of the uplift of the Eastern Cordillera took place relatively recently, during the Pliocene-Holocene (5-3 Mya) (Gregory-Wodzicki, 2000; Hoorn et al., 2010). Simultaneously, the process of closure of the Panama Isthmus connecting North to South America was taking place, creating dynamic connections of land from 30 Mya up to 10 Mya when the full closure was inferred to have occurred (Bacon et al., 2015). The Northern Andes uplift occurred during six phases of activity (Van der Hammen et al., 1973; Zambrano et al., 1971; de González Juana, 1980), but it was not until the maximum period of uplift during the Pliocene (2-5 Mya) that the mountains reached an altitude close to their current height (D-E in Figure 1.3). Nowadays, the Western Cordillera has peaks as high as 4000 m.a.s.l and the Eastern and Central cordilleras have peaks higher than 5000 m.a.s.l.,

a considerable height for *Azteca* species reported at altitudes lower than 1500 m.a.s.l. approximately (Vizek et al., 2012). While the Western and Central cordilleras end in the northern region of Colombia, the Eastern Cordillera continues east to Venezuela as a continuous block of variable heights.

This chapter uses a DNA barcoding approach to assess ant diversity inhabiting *Tococa guianensis* in Colombia, and a phylogenetic approach to assess the temporal match between the diversification of *Tococa*-associated *Azteca* ant lineages and the uplift history of Andean cordilleras. I apply species discovery methods (sensu Collins and Cruickshank, 2013) to sort specimens into MOTUs and address the following questions: *1.* Can any of the sampled MOTUs be identified to species based on high sequence similarity to reference sequences from morphologically identified voucher specimens on NCBI? *2.* How many different ant MOTUs are associated with *T. guianensis* across the sampled areas? *3.* Have *T. guianensis*-associated ants been recorded from other plant hosts? and *4.* Are divergence times of the MOTUs congruent with the uplift of Andean cordilleras? Answering these questions allows me to make a preliminary assessment of the community of ants associated with *T. guianensis*, and to select taxa for more in-depth phylogenomic and population genomic testing of diversification and phylogeographic hypotheses.

2.2 Methods

2.2.1 Sample collection

Collecting sites were selected based on the distribution of *T. guianensis* reported in GBIF (`http://www.gbif.org/species/3858084`, last consulted on March 2016) and

in the Universidad Nacional de Colombia Herbario Virtual (`http://www.biovirtual.unal.edu.co/en/collections/result/species/Tococa%20guianensis/`, last consulted on March 2016). Overall, I visited 17 locations (Figure 2.1,Table B.2 in Appendix B). Additional places with similar forest to that where *T. guianensis* is usually found but where no records of the plant are available were visited, because herbaria records might not reflect the effective distribution of the plant but only areas where collecting efforts have been more. Within each site, sampling took place in at least three areas nearby the collection sites in a radius of 1-5 km to ensure that the diversity present in the area was collected.

Specimens of *T. guianensis* and its ants were collected during two fieldwork expeditions, the first from December 2014 to March 2015, and the second from March to June 2016. Some of the sampling areas originally planned were altered, aborted or unsuccessful for different reasons (Table B.1 with collection codes, plant and ant identities and coordinates are available in Appendix B). The Catatumbo region (Site 1 in Figure 2.1) is of high interest and relevance for its biodiversity, of which only a small proportion has been recorded. Unfortunately, and despite the opening of the region thanks to the recent security improvements, the area is still out of bounds. Unlike Catatumbo, the areas of Tauramena, Amalfi and Cimitarra were formerly restricted areas that now can be accessed by researchers. Collecting samples from those areas increases our knowledge of Colombia's fauna and flora and its distribution in places formerly unexplored. At other areas selected for sampling, the plant was not found (Figure 2.1, Appendix B.2). These regions (along with Antioquia and Santander) have been extensively degraded by deforestation, cattle and agriculture. *T. guianensis* grows in fragmented primary and secondary forest; nonetheless, farmers cut down the plants because of the undesired presence of the ants (according to the locals). In the case of La

Victoria-Caldas, *T. guianensis* records exist at the Universidad Nacional de Colombia Herbarium; however, the area has been flooded to build a dam and I did not find the plant in the surrounding areas. Primary forest in areas like La Victoria and Catatumbo is being lost due to purely economically motivated extraction projects (national and international) before the diversity of the areas can even be recorded. This highlights the relevance of taking conservation and educative actions in rural areas of Colombia, where economic development departs both from conservation and the interest of local people.

2.2.2 Ant sampling, DNA extraction, PCR and sequencing

Azteca colonies were sampled from plants found in a diversity of habitats but commonly growing near water bodies. Worker ants were collected from a minimum of five domatia per plant and placed in collecting tubes with 98% ethanol. When available, alates and larval stages were also collected. To have a glimpse to which ants are associated with *T. guianensis*, the contents of the tubes were first identified morphologically to the genus level using a stereomicroscope and keys to Formicidae (MacKay and Vinson, 1989; Hölldobler and Wilson, 1990; Bolton and others, 1994). Other *Azteca* ants from *Cecropia* trees that I had collected and sequenced in Colombia for a previous project, were included in the phylogenetic dating analyses (see below). Morphological identifications at the genus level were confirmed by comparing all sequences against the NCBI database using the command line **BLAST v.2.6.0** application **Blast+** and the default minimum e-value of $1e^{-25}$. Previous observations on *Cecropia*-associated *Azteca* suggest that each host is inhabited by a single colony once the plant has reached maturity and after exclusion of other competing ant colonies at the seedling stage (Longino, 1989,

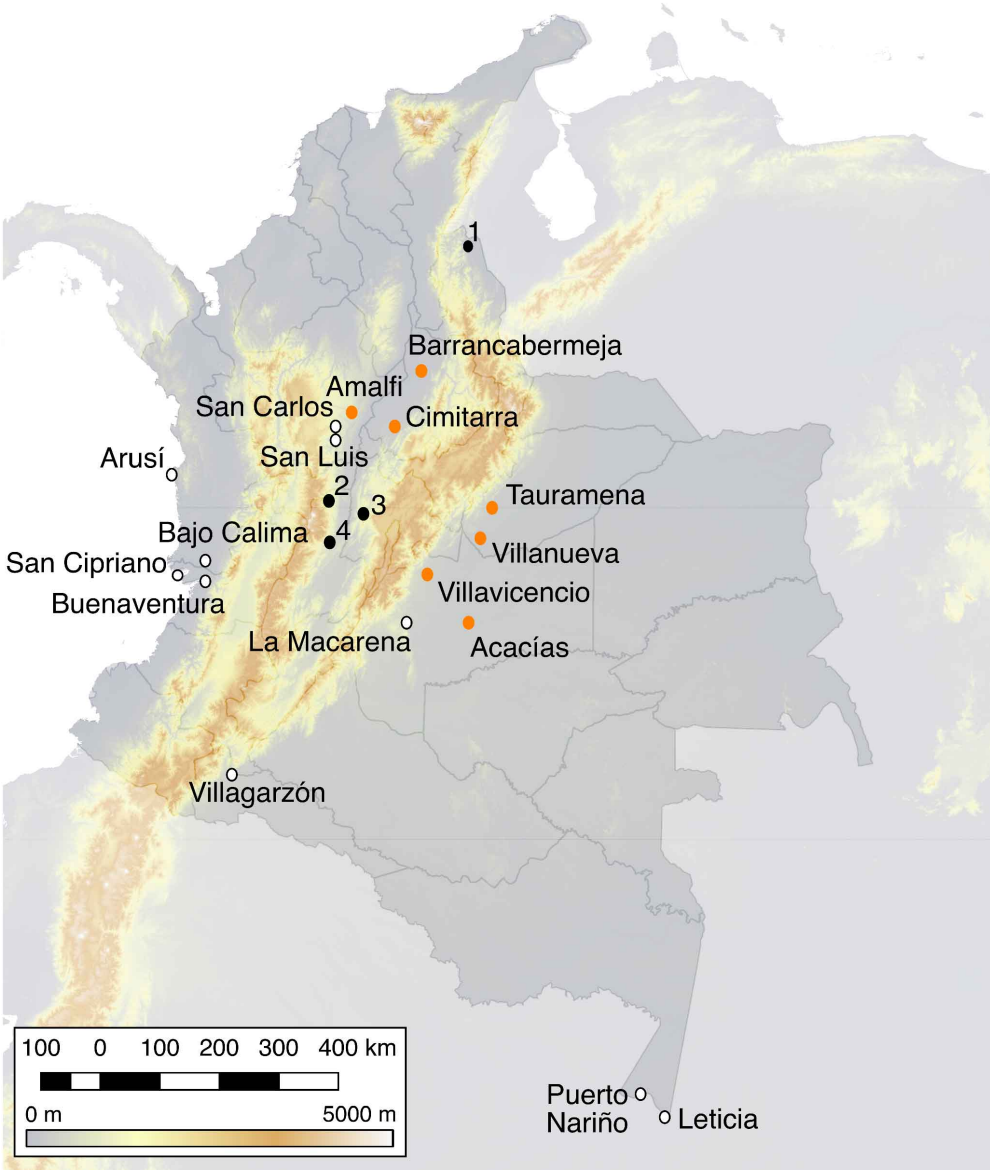


Figure 2.1: Map of Colombia and the location of the collecting sites. White circles represent the first expedition, orange circles represent the second expedition, and black circles represent locations where no *T. guianensis* plants were found or the expedition had to be aborted. Numbers represent locations not included in this study: 1. Catatumbo-Norte de Santander; 2. La Victoria-Caldas; 3. La Dorada-Caldas and Honda-Cundinamarca, 4. Armero-Tolima.

1991c). This has not been studied in *Tococa* plants, but based on fieldwork observations and by collecting only mature plants, a single *Azteca* colony is assumed on each host unless morphologically different ants were collected. Due to limited resources, priority was given to sequence small numbers of ants sampled from different individual host plants rather than many different ants from the same individual host plant. In cases where morphological examination suggested that more than one genus was inhabiting the same host plant individual (10 out of 478 *T. guianensis* specimens), only the *Azteca* ants were sequenced.

One worker ant was selected from each tube, its head removed and the rest of its body crushed with a pestle and placed on a plate wheel. 40μL of 5% Chelex and 5μL of 10mg/mL Proteinase K were added to each wheel and then left incubating overnight in a water bath at 37°C. Finally, plates were centrifuged and heated at 95°C for 15 minutes to denature any remaining enzyme. Regions of the ribosomal internal transcribed spacer region 2 (ITS2) and Cytochrome Oxidase 1 (COI) were amplified using the polymerase chain reaction (PCR). ITS2 is a common nuclear marker used for phylogenetics (Wild, 2009; Hung et al., 2004; Smith et al., 2014) and population genetics (Pringle et al., 2012; Okita and Tsuchida, 2016) and it has been previously used for *Azteca* accessions in reference databases. It is also recommended as a complement to COI for species identification in animals, with a success rate of 91.7% species correctly identified when compared to taxonomy (Yao et al., 2010). The sections at the 5' end and the partial 5.8S sequence are relatively conserved and less variable within populations, making it useful for species-level phylogenetics, while the 3' end tends to accumulate repetitive motifs, indels and inversions, and are used to explore population-level variation (Pringle et al., 2012). Similarly, COI is universally used as a barcode marker for animal species-level phylogenetics and is often also informative at the population level within species

(Hebert et al., 2003a, 2004b; Ward and Holmes, 2007; Linares et al., 2009; Nwani et al., 2011).

The primers used to amplify CO1 were LepF: **5'-ATT CAA CCA ATC ATA AAG ATA TTG G-3'** and LepR: **5'-TAA ACT TCT GGA TGT CCA AAA AAT CA-3'** (Hebert et al., 2004a) and the primers to amplify ITS2 were AW58F1 **5'-AAC GAT TAC CCT GAA CGG TGG A-3**' and AW28S1 **5'-CTG TTC GCT CGC CGC TAC TAA G-3'** (Pringle et al., 2012). For PCR, 1μL of template DNA was added to a final volume of 20μL containing 0.2mM dNTPs mix, 1x PCR Buffer, 2.25mM MgCl2, 0.3μM of each primer, 5μg/μL BSA and 0.3 units of Taq (Bioline). Cycling conditions for COI were 2 min at 94°C followed by 5 cycles of 30 sec at 94°C, 30 sec at 45°C, 40 sec at 72°C; followed by 35 cycles of 30 sec at 94°C, 30 sec at 51°C, 40 sec 72°C, finalizing with 5 min at 72°C. Cycling conditions for ITS2 were 2 min at 94°C followed by 34 cycles of 30 sec at 94°C, 40 sec at 50°C, 1 min at 72°C and finalizing with 5 min at 72°C. PCR products were visualized in a 2% agarose gel stained with SYBRGreen and then cleaned following the shrimp alkaline phosphatase and exonuclease I protocol and subsequently sequenced in both directions on an ABI 3730 capillary machine using BigDye version 3.1 terminator chemistry (Applied Biosystems).

2.2.3 Sequence alignment and phylogenetic inference

Because nDNA and mtDNA have different patterns of inheritance and coalescent history, COI and ITS2 sequences were analyzed separately. Sequences were aligned using the **MUSCLE** algorithm (Edgar, 2004) as implemented in **Geneious v.4.8.5** (http://www.geneious.com, Kearse et al., 2012), then quality checked and edited by

hand. Codon translation was inspected using the insect mitochondrial code and samples whose COI Sequences showed evidence of being a nuclear pseudogene (NUMT), such as unexpected stop codons, shorter sequences, and double bands appearing on the gel, were removed from the analyses. Similarly, a section of approximately 200 bases at the 3' end of the ITS2 alignment was removed as it had multiple indels and repetitive motifs that were difficult to align. Variation in sequences is desirable for species-level studies, but this region in the ITS2 was removed as the homology of the indels and repetitive regions was uncertain. To avoid potential bias introduced by the presence of indels, I generated two sets of sequences per locus, a large set for tree-based identification and a small set for software-based identification. The large sample set for each locus were used for phylogenetic inference and include sequences from ants collected in *Tococa* and *Cecropia* plants in Colombia, *Azteca* sequences available on NCBI, and sequences from the closest available sister genera as outgroups, *Forelius* and *Dorymyrmex* for COI and *Iridomyrmex* and *Linepithema* for ITS2. Outgroups are different for each locus because COI and ITS2 sequences from the same genus are not available on NCBI. Including other species in the phylogenetic reconstructions increases the chances that a query sample will match a conspecific and improves tree topology estimation (Collins et al., 2012; Will et al., 2005).

The small sample sets for each locus only include *Azteca* sequences from specimens collected on *T. guianensis* from which both COI and ITS2 was sequenced successfully, and are used for species delimitation analyses. The reason is that including sequences from distantly related species in the ITS2 alignment resulted in indel variation within *Azteca* clades and populations, and the homology for those indels could not be assessed with certainty. Such variation might cause overestimation of intraspecific distances and subsequently obscure a barcoding gap, misleading the sorting of specimens into MOTUs

(Nicholls, per. comm.). **PartitionFinder v.2** (Lanfear et al., 2016) was used to evaluate appropriate codon partitions based on the Bayesian Information Criterion (BIC) but showed no significant support for any partition scheme over a no partition scheme. A GTR+I+G substitution model was selected for ITS2 and COI using **jModelTest2** based on the above-mentioned BIC (Darriba et al., 2012). For visualizing results and assessing MOTU monophyly, a Bayesian phylogenetic reconstruction for each dataset was performed using **BEAST v.1.8.4** (Drummond et al., 2012) under the GTR+G substitution model, a strict clock model with a lognormal prior distribution, a Birth-Death process (Heled and Drummond, 2015) and a chain of 300 million states length. Bayesian phylogenetic reconstructions are preferred over neighbor-joining clustering methods for sample identification purposes as NJ trees can be misleading if sampling is not complete (Meier et al., 2006; Virgilio et al., 2010; Little, 2011; Zhang et al., 2012; Collins and Cruickshank, 2013). A majority rule consensus tree was obtained using **TREEANNOTATOR v.1.8.2.** (Beast packages, Drummond et al., 2012). Phylogenetic reconstructions using combined COI and ITS2 alignments did not converge (data not shown).

Tree-based identification of specimens followed the criteria established in Meier et al. (2006): a sample is considered successfully identified if the sequence falls in a monophyletic polytomy or clade containing exclusively conspecific sequences. If the sequence falls as a sister to a group of conspecifics, has no conspecific sequences, or forms a polytomy with allospecific sequences, the identification is unsuccessful or ambiguous.

2.2.4 MOTU delimitation

Sequences in the small datasets were clustered into MOTUs using three approaches: **jMOTU** (Jones et al., 2011) and **ABGD** (Puillandre et al., 2012) based on pairwise distances, and **BPP** based on the multispecies coalescent model (Yang, 2015).

(a) **jMOTU** uses pairwise alignments to first group identical sequences and then an all-against-all alignment to generate a distance matrix and cluster sequences depending on their similarities while using different user-provided cut-offs as a reference. It is then up to the user to assess the monophyly of the MOTUs and potential ambiguity of the sample's membership. Because **jMOTU** is provided with a fasta file of unaligned sequences, this approach is more robust to the presence of indels and gaps. Several independent analyses per locus were run to ensure consistency in the results and to evaluate the effect of different combinations of minimum overlap and Megablast identity values. Each scheme generated clusters using a cut-off value ranging from 1 to 100 bases and used values for the minimum overlap and Megablast identity filter parameters between 60% to 95% (and increasing by 5%), for a total of 65 different runs.

(b) **ABGD** takes a file of aligned sequences and a user-provided upper limit that the program uses as the first cut-off to split sequences into MOTUs. The assumption of membership is that sequences from different MOTUs must differ by a higher value than the upper limit, while sequences from the same MOTU must differ from conspecific sequences by a value lower than the upper limit. Once the first set of MOTUs is defined, the algorithm recursively repeats the search within each MOTU until no further gap can be inferred and no further division is possible (Puillandre et al., 2012). To avoid overestimating the number of MOTUs due to potential within-species indel variation in the ITS2 alignment, sequences were separated and aligned into different partitions

based on the clades obtained in the ITS2 phylogeny (Figure 2.5). As the groups are based on discrete clades, the clade is assumed to be the most inclusive possible partition of conspecifics. If the specimens do not belong to the same MOTU, we expect **ABGD** to infer more than one MOTU or one MOTU and related singletons in the clade. If all the specimens included in the partition belong to the same MOTU, only one MOTU is expected. Partitioning the ITS2 alignment reduces the noise introduced by a high number of gaps and indels whose homology is difficult to assess. Moreover, **ABGD** can detect smaller clusters but cannot merge clusters. If a subgroup represents a single clade, further subgroups can be discarded, but the membership of that subgroup to the same clade of another subgroup cannot be assumed. **ABGD** works better when species are represented by more than three or five sequences in the alignment (Puillandre et al., 2012), however, it is less accurate if most potential species are singletons as an estimation of intraspecific distances is not possible from the data. Five independent runs were made for three models of sequence divergence –Jukes-Cantor (JC69), Kimura (K80 with a transition to transversion ratio of 2.0) and simple p-distances–were performed using the online version of ABGD (`http://wwwabi.snv.jussieu.fr/public/abgd/abgdweb.html`), with initial values of Pmin= 0.001 and Pmax= 0.1. The advantage of **ABGD** is the optimization of the threshold from the data instead of relying only on assumptions on the data's membership and the intra and interspecific thresholds (Meyer and Paulay, 2005; Virgilio et al., 2012). MOTUs proposed by **jMOTU** and **ABGD** were validated and accepted if the MOTUs were discrete and monophyletic and if sample membership was unambiguous. MOTUs showing evidence of paraphyly or polyphyly were rejected. Finally, minimum and average intra and interspecific pairwise distances for the main MOTUs delimited using the COI and ITS2 small datasets were calculated using **MEGA v.7.0.26** (Kumar et al., 2016).

(c) **BPP** jointly estimates species delimitation and tree topology under the multi-species coalescent model MSC (Yang, 2002; Rannala and Yang, 2003). The method accounts for present and ancestral coalescent processes involving the populations and species present in the dataset. **BPP** uses a reversible jump Bayesian Chain Monte Carlo (rjMCMC) algorithm to evaluate different species delimitations and the nearest neighbor interchange (NNI) algorithm to evaluate tree topology. This allows the program to account for the uncertainty on the gene tree and performs well even when the information content of the loci is weak and very few loci are used (Yang, 2015). Each sample is assigned *a priori* to a population that the algorithm can merge or keep separated as different MOTUs, but which the program cannot split. After inferring an initial species tree, **BPP** then proposes and evaluates different species delimitation models and estimates the posterior probability distribution of Θ and τ (interpreted as the ancestral effective population size and the root height respectively). **BPP** requires the specimens to be divided *a priori* into groups that can be as small as populations or the smallest monophyletic clades without necessarily reflecting species. To run **BPP** and obtain the posterior probabilities of all delimited MOTUs and because **BPP** can join -but not split- groups, the *a priori* sample groups used in the analysis correspond to the 18 different (and small in terms of sequences included) MOTUs supported by both the COI and ITS2 phylogenies (Figure 6). Splitting the datasets into the smallest possible initial populations prevents **BPP** from underestimating the final number of MOTUs. Because the COI and ITS2 topologies are inconsistent (see Results), analyses were carried out first using both loci and then using one locus at a time. I used the unguided species delimitation algorithm "A11" to calculate the probability of the number of resulting MOTUs and the posterior probability of each group delimited by the algorithm. Priors for Θ were set based on the Theta-W per site estimated for the

small datasets (0.05 and 0.04 for COI and ITS2 respectively) using **DnaSP v.5.10.1** (Librado and Rozas, 2009) and set to 0.045, (gamma distribution with alpha= 1 and beta= 22, or G(1,22)). The priors for τ were set to 0.001 (G(1,1000)) because it represents a more or less recent time to the most recent common ancestor for all *Azteca*. The distribution of the prior is set wide to account for the uncertainty on the ancestral population sizes and time to divergence estimations as recommended in (Yang, 2015). The prior for Θ can be interpreted as the parameter for all population sizes in the data set and the prior for τ as the parameter for divergence time of the root of the species tree. A higher Θ assumes a big population size and a higher τ assumes a longer time to the first divergence time between the pair of lineages. Two independent MCMC runs with a nsample= 20,000 and sampfreq= 5 were carried out for each analysis to check for convergence between runs. The method implemented in **BPP** has been shown to be robust to deviations from the model assumptions, using fairly low numbers of samples and loci (Moritz et al., 2016).

2.2.5 Fossil calibration and Phylogeography

A fossil calibrated phylogeny was reconstructed for each large dataset using **BEAST v.1.8.4** (Drummond et al., 2012). Targeted sampling of plant-ant species is difficult as the identity of the ants remains unknown until the plant is sampled. In addition, the commonest ant species associating with *T. guianensis* will be overrepresented, and occasional ant inhabitants will be singletons. This results in incomplete species sampling overall, with thorough sampling for only a few species. Thus, two models were used for the tree calibration: the birth-death process model (Gernhard, 2008) and the coalescence with constant population size model (Kingman, 1982). Depending on the

species sampling and the samples per species, the use of one or another model is advised. Following Pringle et al. (2012), an uncorrelated relaxed clock model (Drummond et al., 2006) with a lognormal prior with a mean of 0.01 and standard deviation of 0.33 was used for all phylogenetic reconstructions. In addition, a strict clock model was used for the ITS2 dataset as preliminary analyses indicated that chains were not finding the optimum as efficiently when the UCLN model was employed. COI Sequences of *Forelius* and *Dorymyrmex*, and ITS2 sequences of *Iridomyrmex* and *Linepithema* were used as outgroups. Two independent MCMC chains of 300 million generations were run, logging parameters every 3000 generations. Additional runs without alignments were carried out to confirm that priors were not driving the posterior probabilities (Sanders and Lee, 2007). Log files and effective sample size for all parameters were evaluated using **TRACER v.1.8.2** (Beast packages, Drummond et al., 2012). **LOGCOMBINER v.1.8.2** and **TREEANNOTATOR v.1.8.2** (Beast packages, Drummond et al., 2012) were used to combine log and tree files from the three runs, applying a burn-in of 10% of the total number of states. All tree visualizations were done using Baltic (available at `https://github.com/blab/baltic`)

In my analyses, the date of the *Azteca* node was calibrated using a fossil from Dominican amber with an estimated age of between 15-20 Mya during the Miocene. The position of the *Azteca* fossil with respect to the stem or the crown of the genus phylogeny is uncertain as the fossil has not been assigned to a species; however, the fossil sets the limit to how young genus can be. Priors were therefore set with an exponential distribution with an offset of 15 or 20 Mya (to account for the uncertainty of the fossil age) and a mean of 14, such that 95% of the posterior probability distribution of the time to the most recent common ancestor to all *Azteca* includes the age of the stem *Azteca* estimated by Ward et al. (2010) (mean age around 40-45 Mya).

The ancestral areas for the nodes of the *T. guianensis*-associated *Azteca* phylogeny were reconstructed using the Lagrange analysis implemented in **RASP** (Yu et al., 2015) using the ultrametric ITS2 tree as input and removing any accession that is not a *T. guianensis*-associated *Azteca*. Based on the times of Andean uplift and assuming that *Azteca* cannot survive above the 2,000 m.a.s.l., three matrices of dispersal constraints were set as follow: *(i)* from 0 to 3 Mya the probability of migrating from and to areas in the same side of the Andes, *e.g.* A and C, is 1.0 while the probability of migrating to and from areas in opposite sides, *e.g.* A and B, is 0.0; *(ii)* from 3 to 14 Mya, the probability of migrating from and to areas in the same side of the Andes, *e.g.* A and C, is 1.0 while the probability of migrating to and from areas in opposite sides, *e.g.* A and B, is 0.5; *(iii)* from 14 Mya, migration to and from any area has a probability of 1.0. No range constraints were set as there is not enough evidence supporting the absence of ant lineages at an area, but ancestral ranges that do not make biological sense or those including distant areas and excluding areas in the middle were excluded (*e.g.* an ancestral area including A and B but excluding D). Finally, a maximum of 8 areas were allowed for the analysis.

2.3 Results

2.4 Sample collection

The number of ants collected from *T. guianensis* is listed in Appendix B.2 and only in four cases (not listed) out of 420 mature *T. guianensis* had no evidence of inhabitant ants, either because the ant colony died, abandoned the tree or simply never succeeded to colonize the host. Otherwise, all plants were inhabited by ants and their presence was

observed in early stages, even when the plant had only two or three domatia developed. *Azteca* and *Pheidole* were the most common ant genera inhabiting *T. guianensis* and both exhibit different life habits and behavior: *Pheidole* ants build carton nests connected to the domatia entrances throughout the plant stem and their worker behavior is less aggressive than in *Azteca* workers, which additionally do not build carton nests. *Pheidole* was also observed to produce additional entrances to the domatia. Interestingly, *Pheidole* was the dominant inhabitant of *T. guianensis* in Chocó, occupying 61 host plants but was rare in areas like Meta and absent from areas like Valle del Cauca (Figure 2.2, Appendix B.2).

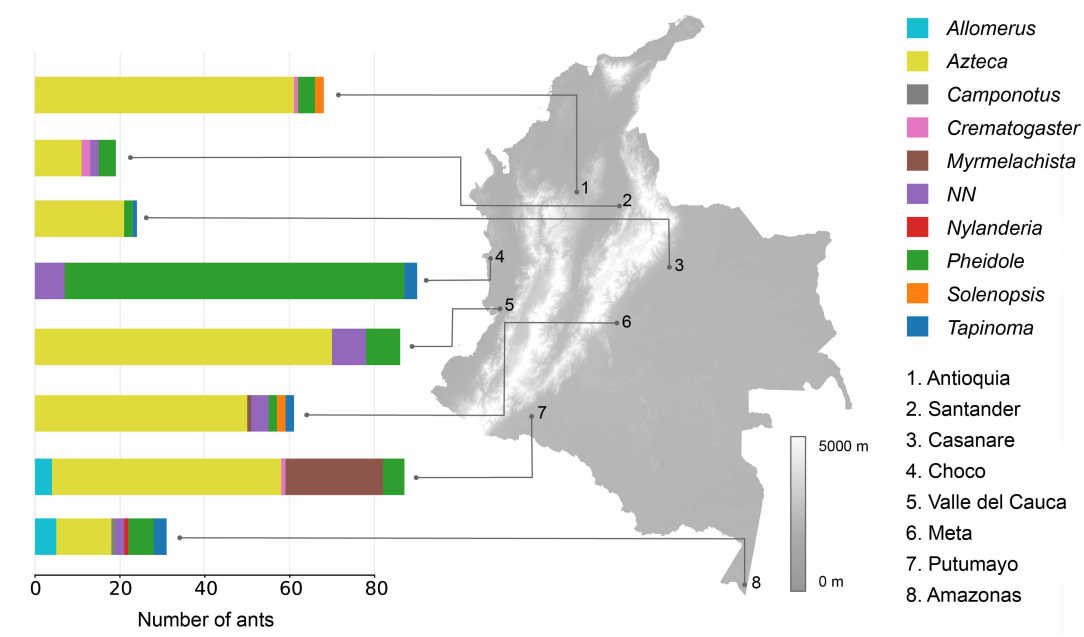


Figure 2.2: Number of ants of each genus collected in Colombia. NN corresponds to unidentified ants.

The large COI dataset includes 335 COI sequences with an amplicon length of 659 bp, in addition to 37 NCBI sequences and 39 *Cecropia*-associated *Azteca* sequences. The large ITS2 dataset includes 218 sequences with sequences from 597 to 986 bp long without indels (which can be more than thousand bp long in reference *Azteca*

accessions), in addition to 34 sequences from NCBI (Table B.4 in Appendix B) and 38 *Cecropia*-associated *Azteca* sequences. The final COI and ITS2 small datasets include 196 *Azteca* sequences, after removing sequences with unexpected stop codons and those not represented in both databases. Nucleotide diversity is higher in COI than in ITS2, and the high Watterson's Theta estimated for the large ITS2 dataset can be attributable to the presence of indels (Table 2.1). In their population study of *A. pittieri*, Pringle et al. (2012) report a similar pattern of rapid evolution in the COI sequences, with an average Pi of 0.00834 for COI and 0.00229 for ITS2.

Based on Blast results, only the hits with the smallest e-value, highest bitscores and identity percentage higher than 98% for species and 80% for a genus, were considered as confirmation of the specimens' identification. Results from blasting the COI sequences identify *Azteca* and *Pheidole* as the most common ants associated to *Tococa*, followed by *Solenopsis* (Myrmicinae), *Tapinoma* (Dolichoderinae) and other rarely collected ants (Figure 2.3). Within *Azteca*, most COI sequences were identified as *Azteca* sp. MAS005 voucher conspecifics, followed by *A. pittieri* and *A. ovaticeps*. However, species-level identification resulted in less than 95% identity between COI query and reference sequences suggesting that *T. guianensis-Azteca* species are not represented in the database (c in Figure 2.3). Low percentages of identity resulted from comparing ITS2 query and reference sequences. Most sequences were identified as *A. beltii* conspecific, but the average percentage of identity is lower than %90 (Figure 2.4). Table 2.2 shows the distribution of e-values, bitscores and percentage of identity for the best hits using COI and ITS2 sequences, with ITS2 resulting in higher identity percentages than COI. Nevertheless, most specimens hit *A. beltii* reference sequences, followed by *A. pittieri*, *A. ovaticeps* and *A. nigricans.* Species identification based on best hits corresponds to the closest *Azteca* reference species available in the NCBI database, but

it is not equivalent to a true identification. At the genus level, Blast results confirmed the morphological identification of most accessions, except for a few *Myrmelachista* specimens morphologically misidentified as *Azteca*.

Table 2.1: Nuclear diversity, length and segregating sites of the mitochondrial COI and nuclear ITS2 sequence alignments.

	Dataset	**Fragment size**	**Segregating sites**	**Pi**	**Theta**
COI	Large	659	294	0.12	48.79
	Small	659	250	0.1	42.71
ITS2	Large	597-986*	625	0.05	103.72**
	Small	905	208	0.02	35.54

*Without indels

* Calculated from the alignments with indels.

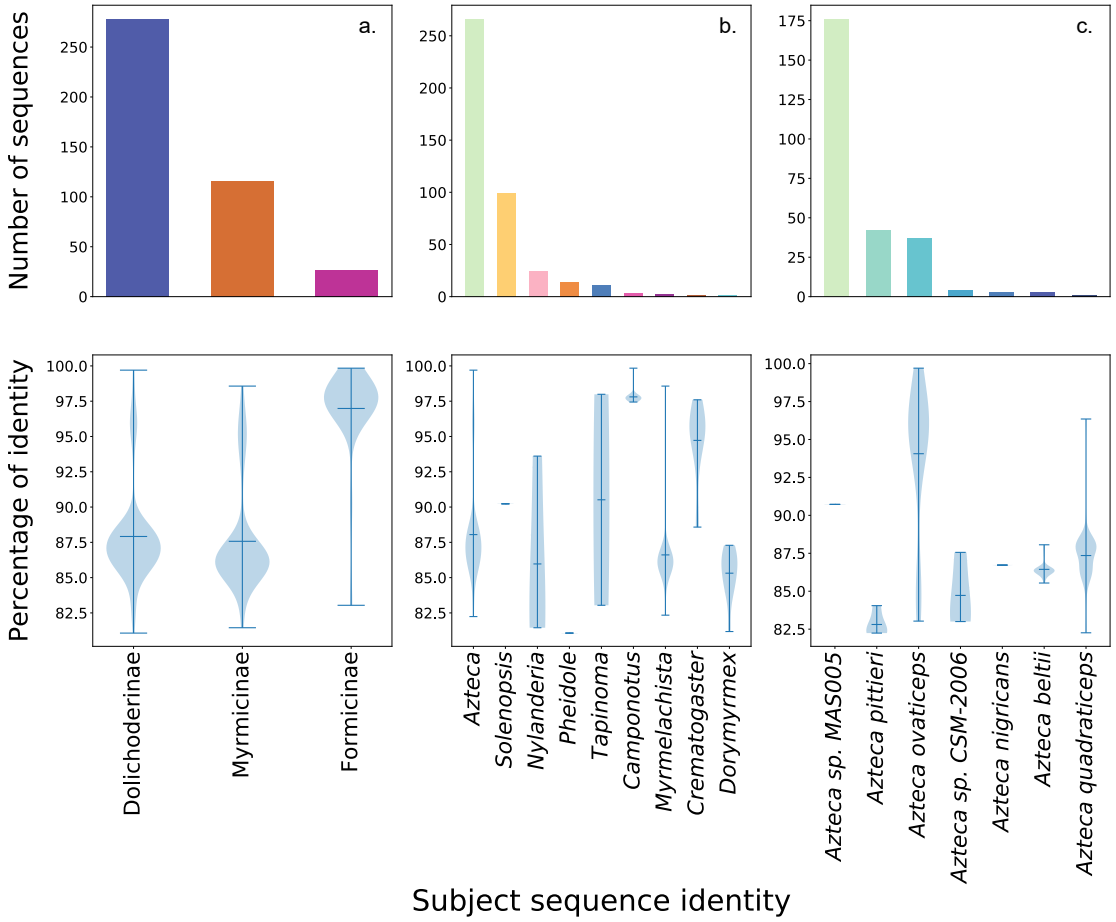


Figure 2.3: Identity of COI sequences from *T. guianensis*-associated ant specimens. Sequence identification based on the best hit against NCBI subject sequences is shown at the top and the percentage of identity between query and subject sequences is shown at the bottom. **a.** Identification to the subfamily level. **b.** Identification to the genus level. **c.** Identification to the species level.

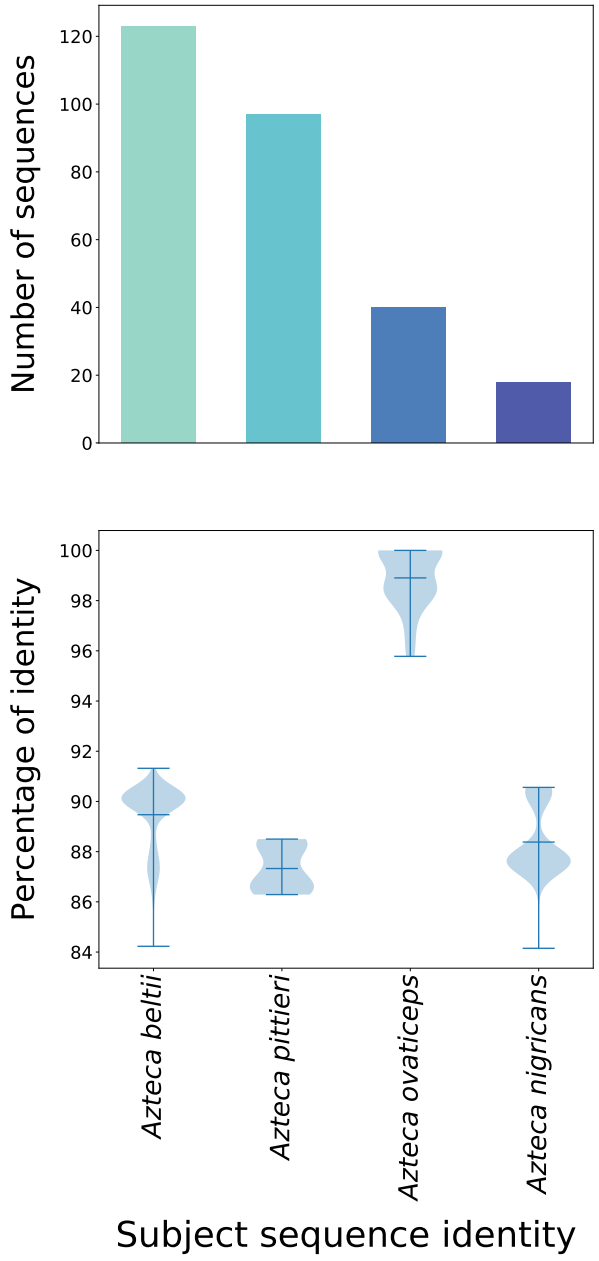


Figure 2.4: Identity of ITS2 sequences from *T. guianensis*-associated ant specimens. Sequence identification based on the best hit against NCBI subject sequences is shown at the top and the percentage of identity between query and subject sequences is shown at the bottom.

Table 2.2: Cut-off points of the distribution of e-values, bitscores and identity percentage for the best Blast hits of COI and ITS2 sequences.

Marker	**Score**	**Lowest value**	**First quantile**	**Mean**	**Third quantile**	**Highest value**	**Median**
	E-value	0	0	1.44E-143	0	6.00E-168	0
COI	Bitscore	512	695	772.90974	784	1205	712
	Identity percentage	81.07	86.15	88.402945	88.24	99.84	86.61
	E-value	0	0	3.59E-35	0	1.00E-32	0
ITS2	Bitscore	152	880	913.44	996	1415	976
	Identity percentage	84.15	87.65	90.31	90.15	100	90.15

2.5 Sequence alignment and phylogenetic inference

The final alignments of the large and small datasets are 659 bp and 1354 bp long for COI and ITS2 respectively. In general, COI and ITS2 support the same backbone topology and recovered branches have posterior probabilities ranging from poor (below 0.7), to medium (between 0.7 and 0.9) and good support (above 0.9). Both phylogenies show a split between *Tococa*-associated *Azteca* and other *Azteca* commonly found in other host plants (Figure 2.5 and 2.6). Moreover, both loci recover a clear split between *Azteca* collected from *T. guianensis* to the east of the Eastern Andes Cordillera (from here on referred to as Eastern *Azteca*) and those collected to the North and West of it (from here on referred to as Western *Azteca*). The phylogenetic reconstruction of the large COI dataset (Figure 2.5) revealed a polytomy of four major clades: *Forelius* and *Dorymyrmex* outgroup sequences, two clades of *Cecropia*-associated *Azteca* from Colombia (C) and other plant-associated *Azteca* clustering with the NCBI reference sequences (A), and a clade of *Tococa*-associated *Azteca* (T). The clade T further divides into one Santander subclade (S1) sister to other two grouping *Azteca* subclades: one of Eastern *Azteca* (Casanare, Meta, Putumayo and Amazonas), and the other of Western (Antioquia and Valle del Cauca) and more Santander specimens (S2). A strong geographic structure can be observed within Western *Azteca* as specimens from Valle del Cauca and Antioquia reflect the presence of the Western Cordillera between both populations (Figure 2.5). Less structure is observed within Eastern *Azteca*, while all samples from Putumayo cluster together, samples from Meta and Casanare interleave.

Internal branches on the ITS2 phylogeny have a higher posterior probability compared to COI results, but branch posterior probabilities decrease towards the tips. However, the main clades are consistent with those of COI except for clade C and a few specimens

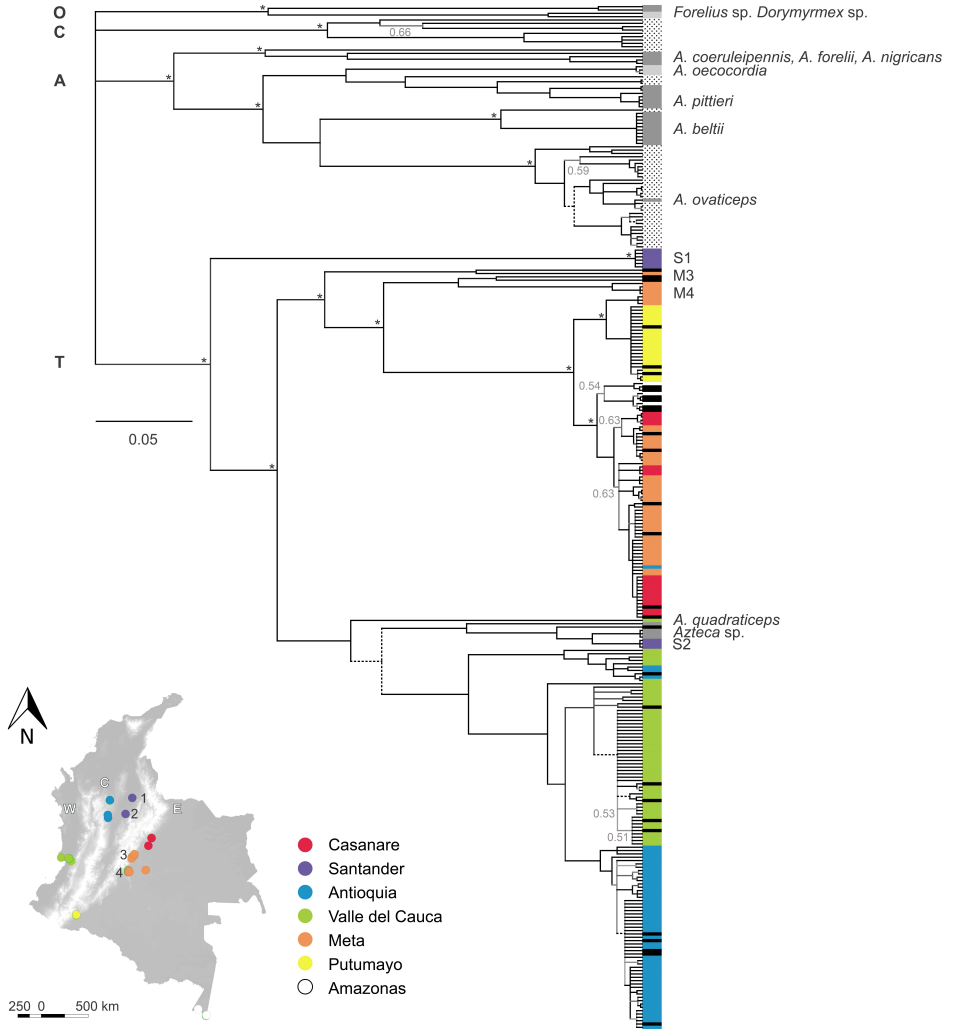


Figure 2.5: Majority consensus tree obtained from the large dataset of *Azteca* COI sequences including sequences from NCBI (in grey), sequences from *Cecropia*-associated *Azteca* collected in Colombia (represented by dotted filling) and *Forelius* sp. and *Dorymyrmex* sp. outgroup sequences. Specimens in black are *Tococa*-associated *Azteca* not included on the species delimitation analyses. Support values correspond to the posterior probability of the branch and only values between 0.5 and 0.7 are shown. Black branches have a support higher than 0.7, branches with less than 0.4 posterior probability are collapsed and dotted branches have a support between 0.4 and 0.6. Asterisks indicate branches with a posterior probability higher than 0.9. The map shows the collecting sites and the colour code for the specimens. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

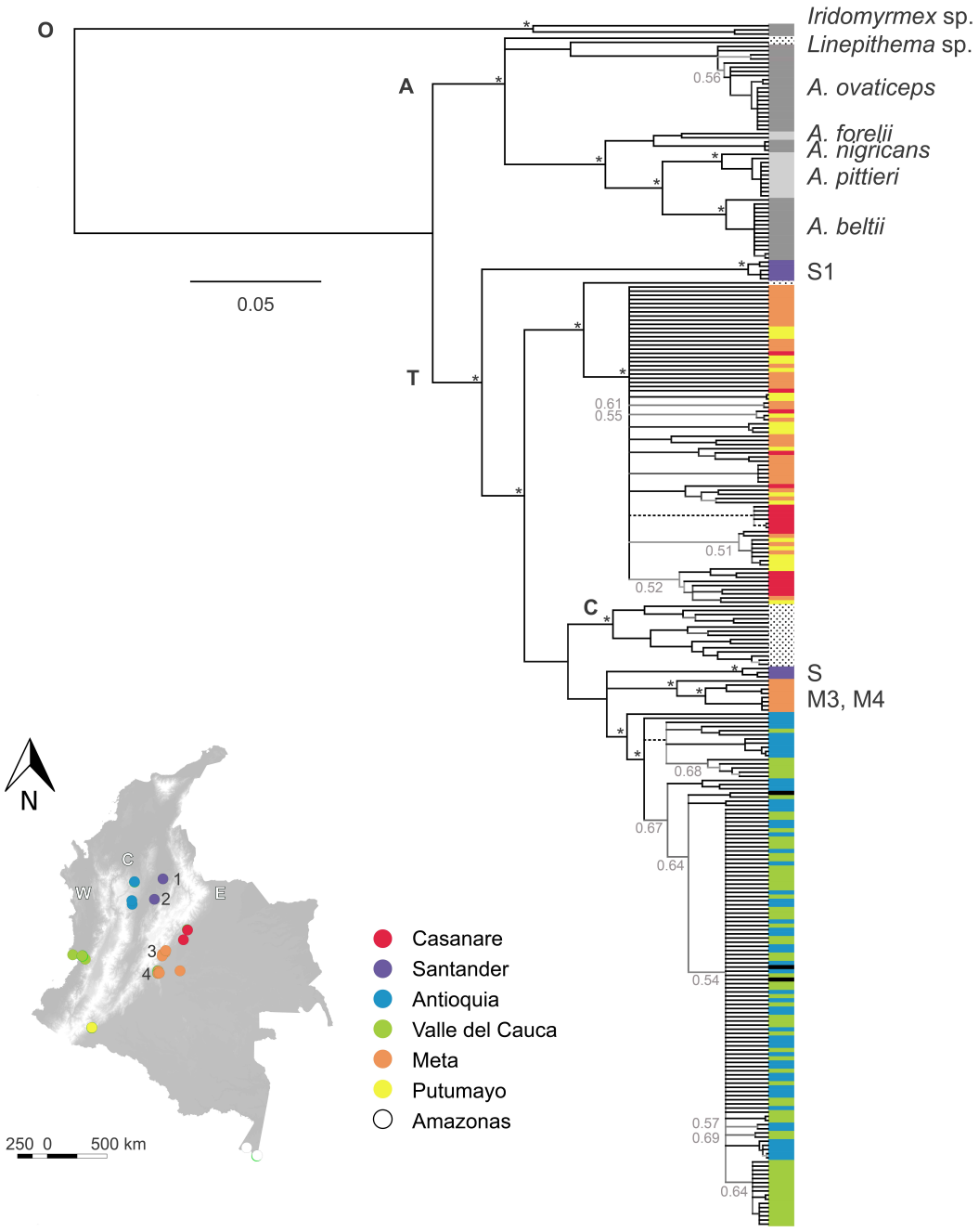


Figure 2.6: Majority consensus tree obtained from the large dataset of *Azteca* ITS2 sequences including sequences from NCBI (in grey), sequences from *Cecropia*-associated *Azteca* collected in Colombia (represented by dotted filling) and *Iridomyrmex* sp. and *Linepithema* sp. outgroup sequences. Specimens in black are *Tococa*-associated *Azteca* not included on the species delimitation analyses. Support values correspond to the posterior probability of the branch and only values between 0.5 and 0.7 are shown. Black branches have a support higher than 0.7, branches with less than 0.4 posterior probability are collapsed and dotted branches have support between 0.4 and 0.6. Asterisks indicate branches with a posterior probability higher than 0.9. The map shows the collecting sites and the color code for the specimens. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

from Meta (M3 and M4 in Figure 2.6). The placement of the outgroup (O) and the *Tococa*-associated *Azteca* clade (T) with respect to the rest of the tree is the same as in the COI phylogeny. Clade T consists of the S1 *Azteca* as sister of two subclades including Western, Eastern and *Cecropia*-associated *Azteca* sampled in Colombia. Geographic structure dividing Eastern and Western *Azteca* is recovered by ITS2 apart from the M3 and M4 groups placed with the rest of Western *Azteca*. The placement of the C clade differs from COI as its position is reconstructed as a sister clade to the rest of Western *Azteca*+S2+M3+M4 within that clade (Figure 2.6). The position of M3 and M4 groups in the phylogenies of the large and small datasets (Figure 2.7) might be conflicting possibly due to either evolutionary processes or a sampling issue additional to not enough information accumulated on the ITS2 marker, as suggested by the poor resolution and branch support. The position of the two groups from Santander (S1 and S2) is recovered consistently using the large datasets from both loci, with the Barrancabermeja group (S1) sister to the rest of the T clade and the Cimitarra group (S2) is more related to the western *Azteca*, but their placement is inconsistent when the small datasets are used. Less geographic structure is observed within Eastern *Azteca* specimens.

Phylogenies from the COI and ITS2 small datasets recovered the same split between Western and Eastern *Azteca* specimens with slight differences in the topology when compared to each other and to the large datasets, mainly in the case of ITS2 (Figure 2.7). First, S2 forms a polytomy with S1 and the rest of Eastern *Azteca* while it is recovered as sister to most Western *Azteca* in the large datasets and the COI small dataset. Second, M3 and M4 are recovered as part of the Eastern *Azteca* and not within the Western *Azteca* as with the ITS2 large dataset. Topologies recovered by either the small or large COI datasets are otherwise the same.

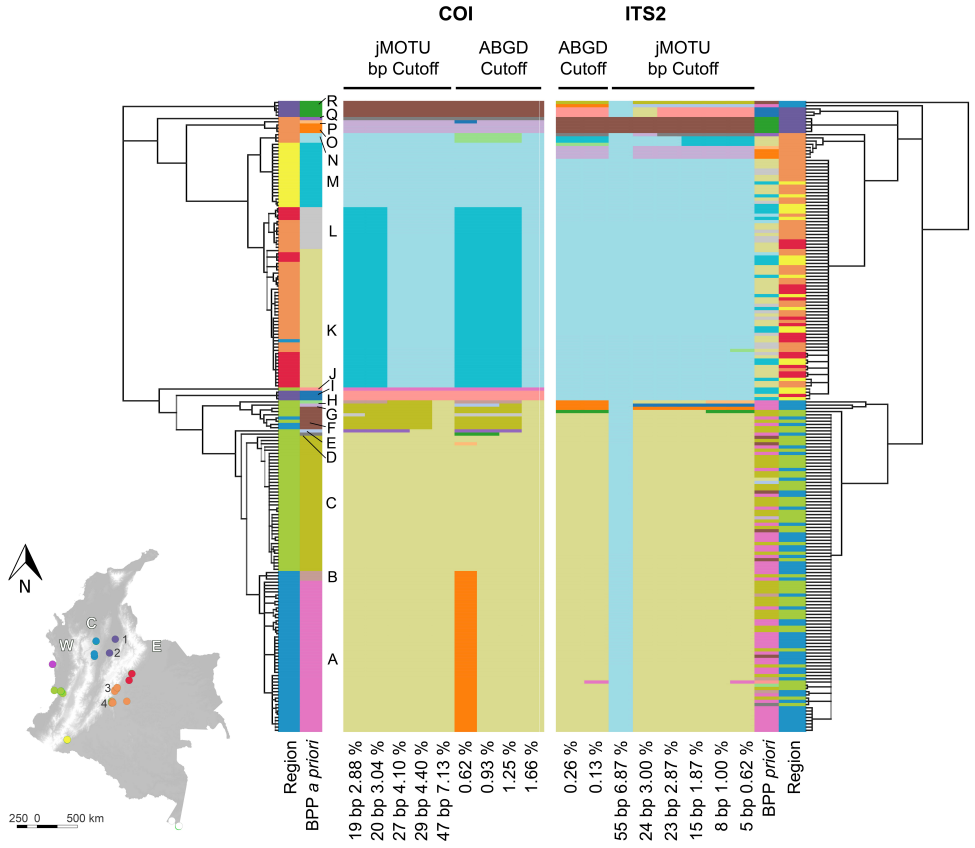


Figure 2.7: MOTUs delimited by **jMOTU** and **ABGD** and a majority consensus rule tree reconstruction of the COI and ITS2 small datasets. Branches with less than 0.40 posterior probability have been collapsed. **BPP** *a priori* letters correspond to the groups used on **BPP** analyses. Tips of both phylogenies are colored according to the regions shown in the map. **jMOTU** cut-off values are reported both in base pairs and in terms of the percentage this represents of the total sequence length. **ABGD** cut-off values represent the JC69 distance.

2.5.1 MOTU delimitation

MOTU delimitation resulted in the majority of COI and ITS2 sequences clustered into two large MOTUs consistently defined by both, **jMOTU** and **ABGD**. These two MOTUs cluster the majority of the *T. guianensis*-associated *Azteca* sequences and are consistent with the Eastern and Western lineages seen in the phylogenies (Figure 2.8 and Figure 2.6). However, both the total number of MOTUs defined and the cut-off values estimated differ slightly among loci and approach, a pattern resulting from the

presence of singleton sequences and the differences in evolutionary rates between a nuclear (ITS2) and mitochondrial (COI) marker (Tables B.5 and B.6 in Appendix B). Delimitation of MOTUs is more robust when groups are represented by a large number of sequences. Moreover, independent **jMOTU** runs varying the minimum overlap and Megablast identity filter parameter values resulted in no difference in the MOTUs delimited. Thus, only the results obtained using a minimum overlap and Megablast identity filter of 95% are shown. Similarly, the use of different genetic distance measures (JC69, K80 or simple p-distances) did not change the MOTUs reported by ABGD, and thus only results using the JC69 distances are shown.

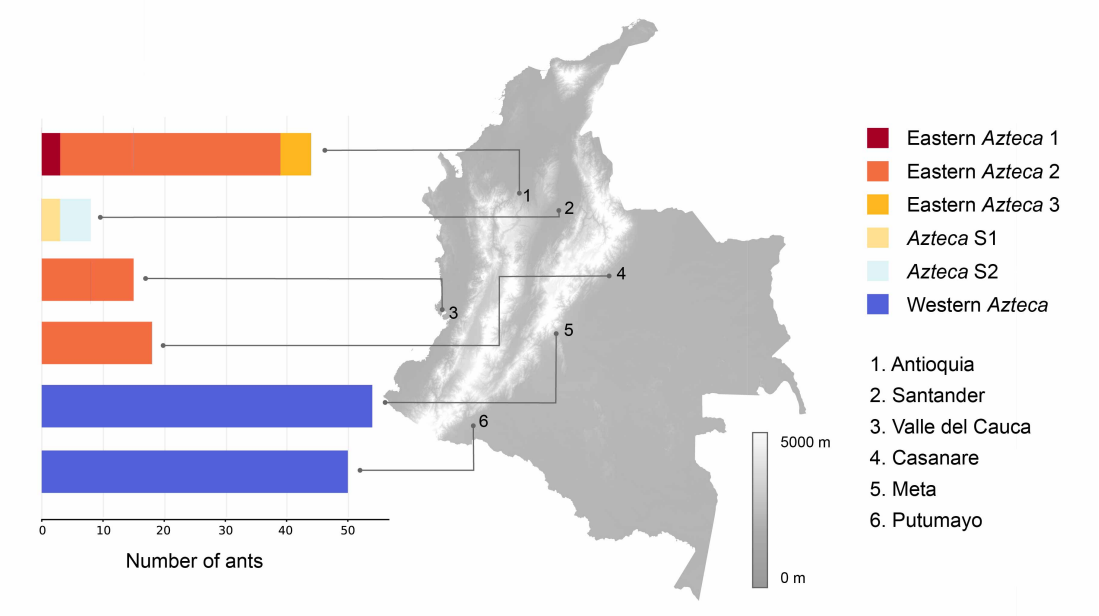


Figure 2.8: Number of ant sequences per lineage and their collecting sites. The lineages include Eastern and Western *Azteca*, the *Azteca* lineages sister to Eastern *Azteca*, and the Santander 1 and 2 lineages (S1 and S2).

Unaligned small datasets used for species delimitation analyses resulted in 659 and an average of 807 bp for COI and ITS2 respectively. **jMOTU** and **ABGD** resulted in similar MOTU configurations for both loci despite the topological differences (Figures 2.9 and 2.10), delimiting a minimum of three and a maximum of 18 different MOTUs

across loci when using different cut-off values. Delimitation of MOTUs based on ITS2 resulted in two large MOTUs corresponding to Western *Azteca* and Eastern *Azteca*. The same MOTUs were delimited based on COI by cut-off values higher than 27 bp or a JC96 distance of 1.66%. Lower cut-off values for COI delimitation resulted in nested groups separating *Azteca* by region. Reconciling geographical and phylogenetic information from both loci with the different delimitation schemes, **ABGD** results suggest the best delimitation threshold for the COI sequences is 1.66% resulting in 5 MOTUs and 2 singletons. **jMOTU** results suggest the best threshold is 47, bp 7.13% of the average total sequence length, resulting in the same five MOTUs and two singletons. For ITS2, the best **ABGD** threshold is 0.26% resulting in seven MOTUs and five singletons, while the best threshold suggested by **jMOTU** is eight bp or 1.0% of the average total sequence length, resulting in five monophyletic MOTUs, one paraphyletic MOTU and seven singletons. Within and between MOTU distances were calculated after reconciling the MOTUs delimited by **jMOTU** and **ABGD** (Table 2.3). Within MOTU distances are consistently smaller than between MOTUs, and distances are the same when calculating using p-distance, JC69 or K2P distances (only K2P distances are shown). Additionally, distances are lower for ITS2 sequences than for COI sequences, as expected given the known higher evolutionary rate of COI (COI Pi values of 0.12-0.10 for the large and small datasets compared to 0.05 and 0.02 for ITS2, Table 2.1). The average intraspecific distance for COI reported for other Hymenoptera is 0.018 (± 0.022), whilst the average smallest interspecific distance is 0.038 (±0.042) and the average interspecific distance is 0.093 (±0.039) (Meier et al., 2008).

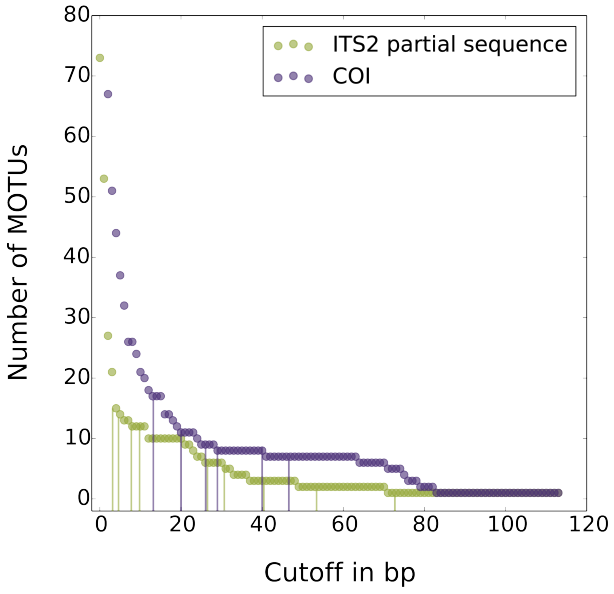


Figure 2.9: Number of MOTUs of the COI and ITS2 sequences delimited by **jMOTU** using cut-off values (in base pairs) from zero to 120.

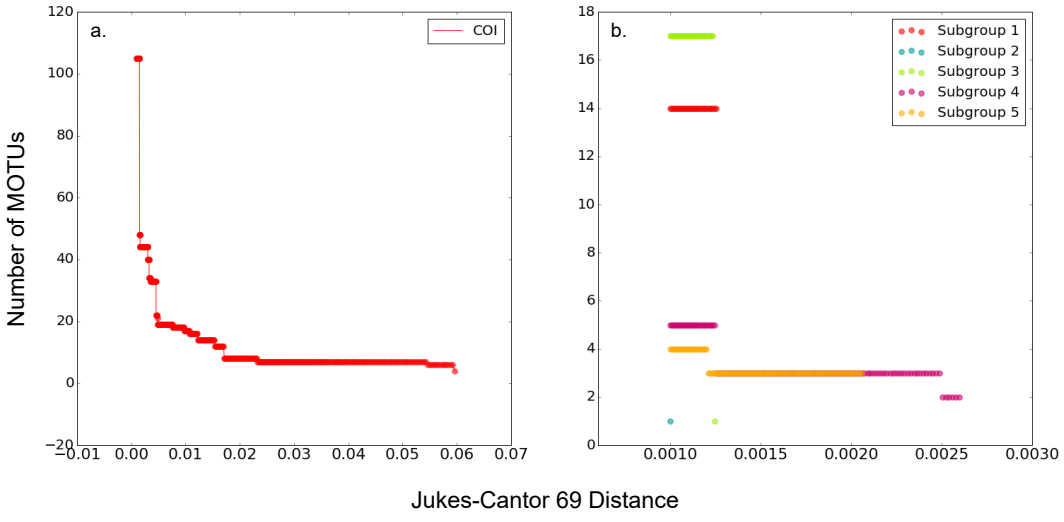


Figure 2.10: Number of MOTUs of the COI and ITS2 sequences delimited by **ABGD** using cut-off values (based on JC69 pairwise distances). **a.** MOTUs of COI sequences. **b.** MOTUs of ITS2 subgroups of sequences.

Table 2.3: Pairwise distances within and between MOTUs.

Marker	MOTU	Average within MOTU	Between MOTU	
			Minimum	Average
COI	Eastern	0.015	0.108	0.123
	Western	0.023	0.096	0.114
	S1	0	0.127	0.133
	S2	0	0.096	0.112
ITS2	Eastern	0.002	0.02	0.038
	Western	0	0.016	0.031
	S1	0	0.063	0.061
	S2	0.003	0.016	0.036

BPP evaluates the probabilities of the number of final delimited MOTUs, from one up to the number of *a priori* groups provided to the program, and assigns this probability to each final number of MOTUs. Analyses using ITS2 and COI resulted in the highest posterior probability value supporting a total of 15 MOTUs (Table 2.4). However, after controlling for prior probabilities, the ratio between posterior and prior probabilities is higher for 18 MOTUs. Table 2.5 shows the posterior probabilities of the *a priori* groups (**BPP** *a priori* groups in Figure 2.7) and the different combinations of joint groups delimited by **BPP**. Overall, posterior probability decreases as the *a priori* groups included within MOTUs increases. Posterior probabilities for most *a priori* groups are higher than 0.8 except for groups formed by singletons. Within Western *Azteca*, groups A, B, C and D were also expected to form a single MOTU, however, **BPP** failed to recover the MOTU and instead recovered each group individually with a relatively

high posterior probability. Similarly, groups K, L, M, N and O, were not recovered by **BPP** as Eastern *Azteca*, although these were expected to belong to a single MOTU based on the ITS2 phylogeny. It is possible that topological conflicts between COI and ITS2 and the strong geographic structure of COI compared to ITS2 are biasing MOTU delimitation towards non-conflicting, *a priori* groups supported by COI.

Table 2.4: Prior and posterior probabilities for each model of MOTUs delimited by **BPP**.

No. MOTUs	**Prior Probability**	**Average posterior probability**	**Posterior/Prior ratio**
1	0.0044	0	0.00
2	0.0044	0	0.00
3	0.0064	0	0.00
4	0.01	0	0.00
5	0.0158	0	0.00
6	0.0247	0	0.00
7	0.0374	0	0.00
8	0.0546	0	0.00
9	0.0758	0.0002	0.00
10	0.0991	0.006	0.06
11	0.1206	0.007	0.06
12	0.1343	0.023	0.17
13	0.1343	0.094	0.70
14	0.1174	0.18	1.53
15	0.0862	0.221	2.56
16	0.05	0.213	4.26

Table 2.4 continued from previous page

17	0.02	0.173	8.65
18	0.0044	0.081	18.41

Table 2.5: Posterior probabilities for each *a priori* group and the MOTU delimited by **BPP**.

Western *Azteca*										**Eastern *Azteca***							
A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**
0.95	0.78	0.83	0.79	0.79	0.77	0.76	0.76	0.98	0.85	0.99	0.99	1	1	0.78	0.68	0.82	0.99
0.001		0.83	0.79	0.79	0.77	0.76	0.76	0.98	0.85	-		1	1	0.78	0.68	0.82	0.99
-				0.79	0.062		0.76	0.98	0.85	-		-		0.21		0.82	0.99
-				0.003				0.98	0.85	-				0.21		0.82	0.99

2.5.2 Fossil calibration and Phylogeography

The three nodes of interest are the split between *Azteca* and its outgroup, the common ancestor for all *Azteca*, and the split between Western and Eastern *Azteca*. For all three, the estimated mean age tends to be older and confidence intervals wider under the coalescent model when compared to the birth-death model (Figure 2.11). Similarly, time estimates for the split between outgroups and *Azteca* are less consistent among tree and clock models than they are for the other two nodes. *Azteca* and the outgroup diverged around 33 and 82 Mya, with ITS2 producing the oldest estimates when the prior offset is set to 20 Mya, the oldest boundary of the fossil age range. The mean time to the most recent common ancestor (tMRCA) to all *Azteca* is between 15 and 32 Mya as the node can be older but not younger than the *Azteca* fossil. Finally, mean time to the divergence between Western and Eastern *Azteca* is estimated between 12 and 25 Mya for both ITS2 and COI, with the oldest estimates obtained from ITS2 (Figures 2.12 and 2.13). Confidence intervals for the three node ages overlap to a large extent suggesting that the information in both loci is not enough to estimate the timing of divergence events with certainty (Table 2.6). This calibration was estimated using a subsample of sequences from each *Azteca* species available in NCBI. However, the results reported here are consistent with an ITS2 calibration of *Azteca* using all sequences available in NCBI, demonstrating that the topological patterns and the time of divergence are not influenced by the sequence sampling included in the analysis (see Appendix B, Figure B.1).

Table 2.6: Mean, median and 95% highest posterior density (HPD) intervals of the tMRCAs for *Azteca* lineages estimated from COI and ITS2 using different tree and clock models. Values inside parentheses correspond to the median. HPD were calculated for phylogenies using the oldest (20 Mya) and youngest (15 Mya) date estimates for the *Azteca* fossil.

		COI		**ITS2**			
		UCLN		**UCLN**		**Strict**	
Node	**Fossil calibration**	**Coalescence**	**Birth death**	**Coalescence**	**Birth death**	**Coalescence**	**Birth death**
All ants	*15*	27.26 (22.53)	25.36 (22.61)	47.44 (39.12)	26.50 (23.41)	66.68 (57.89)	45.79 (40.77)
	95% CI	15.00 - 53.74	15.20 - 43.47	15.07 - 102.99	15.04 - 46.58	33.48 - 125.49	26.7 - 77.64
	20	42.69 (37.25)	33.04 (29.83)	57.42 (48.10)	33.89 (30.48)	82.70 (72.91)	57.54 (52.09)
	95% CI	20.25 - 80.44	20.13 - 54.48	20.45 - 119.30	20.25 - 57.54	45.00-146.58	35.11 - 94.40
Azteca	*15*	23.58 (20.58)	21.64 (19.23)	24.86 (21.53)	22.00 (19.47)	24.72 (21.22)	22.07 (19.61)
	95% CI	15.00-42.06	15.00 - 36.23	15.00 - 44.84	15.0 - 37.39	15.00 - 45.57	15.00 - 36.88
	20	30.86 (27.14)	27.32 (24.63)	30.54 (27.09)	27.98 (25.25)	30.55 (26.83)	27.75 (25.02)
	95% CI	20.00 - 54.17	20.00-43.11	20.00 - 52.25	20.00 - 44.59	20.00-53.53	20.00 - 44.62
Tococa-Azteca	*15*	12.84 (11.53)	16.76 (15.25)	16.34 (14.49)	17.52 (15.72)	21.11 (18.19)	18.82 (16.76)

Table 2.6 continued from previous page

	95% CI	4.12 - 23.24	7.01 - 29.29	4.91 - 32.38	8.08 - 31.32	11.79 - 39.23	11.59 - 32.02
	20	20.47 (18.87)	21.82 (20.25)	20.38 (18.55)	22.15 (20.33)	26.13 (22.91)	23.71 (21.39)
	95% CI	6.32 - 36.31	11.68 - 35.94	6.99 - 38.35	10.66 - 37.55	15.38 - 47.18	15.49 - 39.09
	15	10.64 (9.31)	12.87 (11.76)	9.81 (8.56)	12.96 (11.70)	17.86 (15.34)	16.06 (14.34)
Eastern-Western	*95% CI*	3.16 - 21.62	5.47 - 22.91	2.78 - 20.04	5.33 - 23.29	9.99 - 33.43	9.48 - 27.14
Azteca	*20*	15.52 (13.86)	17.16 (15.92)	12.22 (10.89)	16.35 (15.12)	22.06 (19.42)	20.25 (18.30)
	95% CI	4.93 - 29.60	7.55 - 29.22	3.56 - 23.77	7.05 - 28.33	12.88 - 38.13	12.69 - 32.90
	15	7.15 (6.30)	9.62 (8.73)	7.62 (6.58)	10.84 (9.79)	13.49 (11.62)	12.13 (10.80)
Eastern	*95% CI*	2.05 - 14.36	3.46 - 17.57	2.29 - 15.95	428 - 19.95	6.94 - 24.85	7.28 - 20.87
Azteca	*20*	10.89 ()	12.25 (11.29)	9.52 (8.4)	13.64 (12.58)	16.58 (14.57)	15.31 (13.87)
	95% CI	2.76 - 21.44	4.09 - 22.79	2.66 - 18.80	5.31 - 23.7	9.24 - 29.40	9.10 - 25.37
	15	5.98 (5.24)	5.26 (4.70)	7.39 (6.40)	8.40 (7.58)	10.15 (8.77)	9.39 (8.33)
Western	*95% CI*	1.43 - 12.46	1.72 - 10.05	1.85 - 15.64	2.72 - 16.03	5.12 - 18.67	5.35 - 15.91
Azteca	*20*	6.81 (5.98)	6.71 (6.05)	6.07 (5.24)	10.4 (9.53)	12.76 (11.21)	13.35 (12.12)

Table 2.6 continued from previous page

95% CI	2.49 - 13.20	2.72 - 12.50	1.64 - 12.71	3.01 - 19.07	7.16 - 23.12	6.79 - 19.03

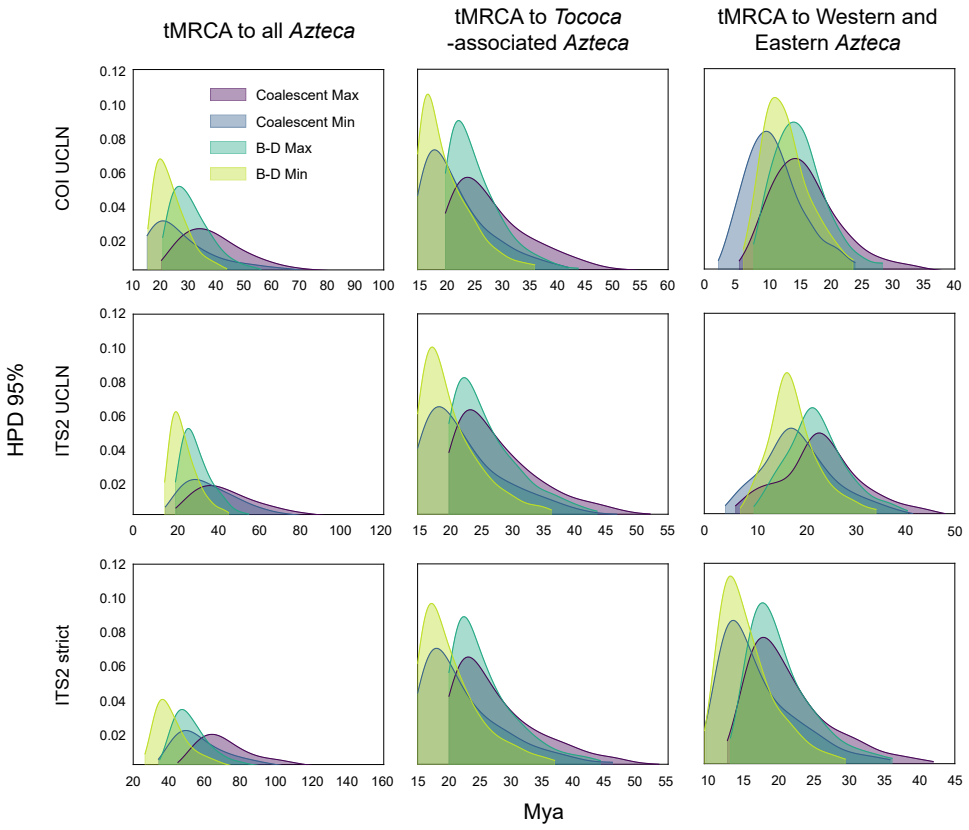


Figure 2.11: 95% highest posterior density interval for the tMRCA to all *Azteca*, *Tococa*-associated *Azteca* and the split between Western and Eastern *Azteca* for different tree and clock models.

Reconstruction of the ancestral areas suggests that Most Recent Common Ancestor (MRCA) to all *T. guianensis*-associated *Azteca* was present in the Northern area of the Andes (Figure 2.14). Probabilities for the node are low, but all results are either Santander or Antioquia, and in one case, Casanare. Ancestral areas for other nodes in the phylogeny are better supported, placing the MRCA to both Eastern and Western *Azteca* in Meta and possibly Casanare. The ancestral area for Western *Azteca* includes Meta and Antioquia, as does the ancestral area of MRCA to S2, M3 and M4. Finally, the ancestral area of the MRCA to Eastern *Azteca* includes Meta and Casanare.

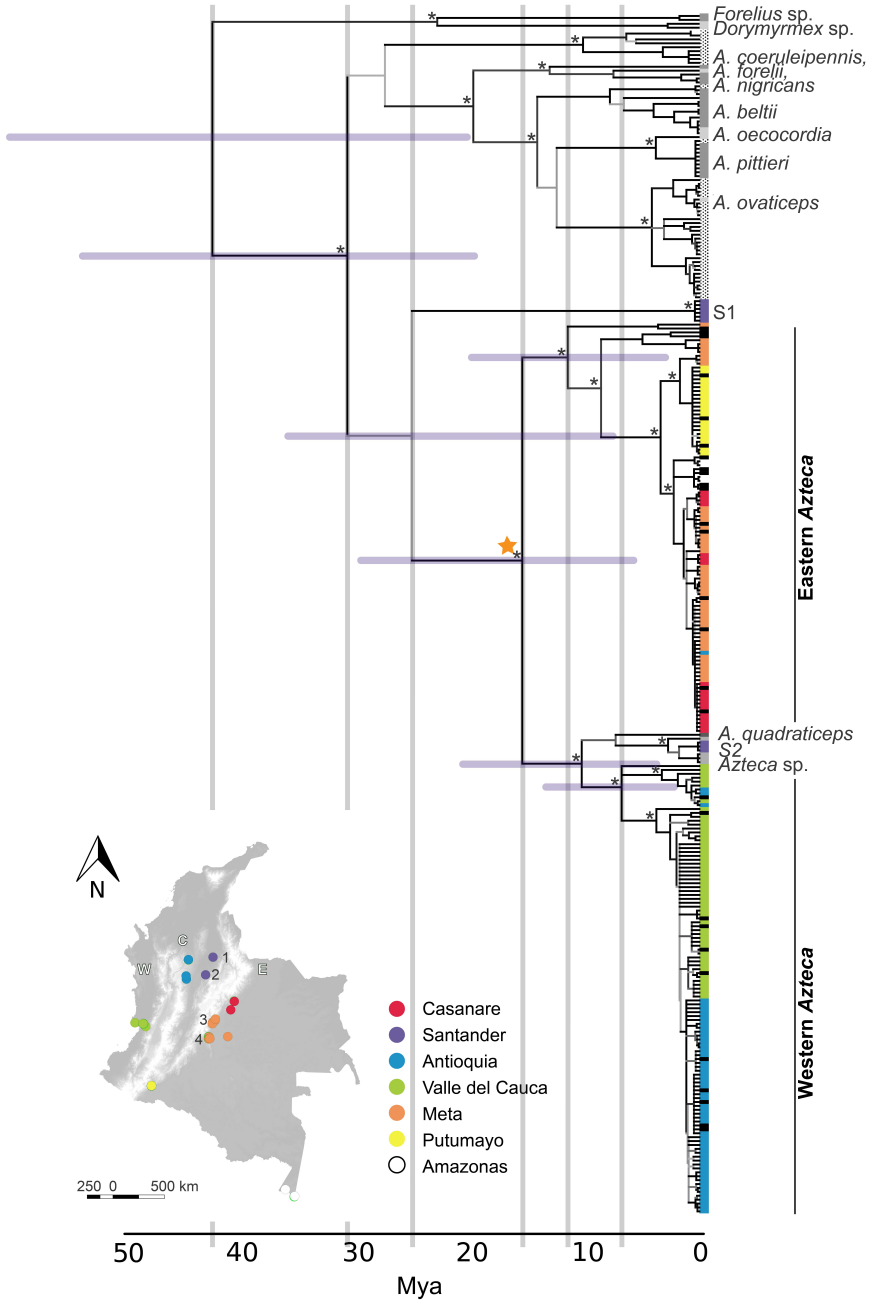


Figure 2.12: Calibrated majority consensus tree obtained from the large dataset of *Azteca* COI sequences including sequences from NCBI (in grey), sequences from *Cecropia*-associated *Azteca* collected in Colombia (represented by dotted filling) and *Forelius* sp. and *Dorymyrmex* sp. outgroup sequences. Specimens in black are *Tococa*-associated *Azteca* not included on the species delimitation analyses. The star indicates the split between Western and Eastern *Azteca*. Support values correspond to the posterior probability of the branch and only values between 0.5 and 0.7 are shown. Black branches have a support higher than 0.7, branches with less than 0.4 posterior probability are collapsed and dotted branches have a support between 0.4 and 0.6. Asterisks indicate branches with a posterior probability higher than 0.9. The map shows the collecting sites and the color code for the specimens. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

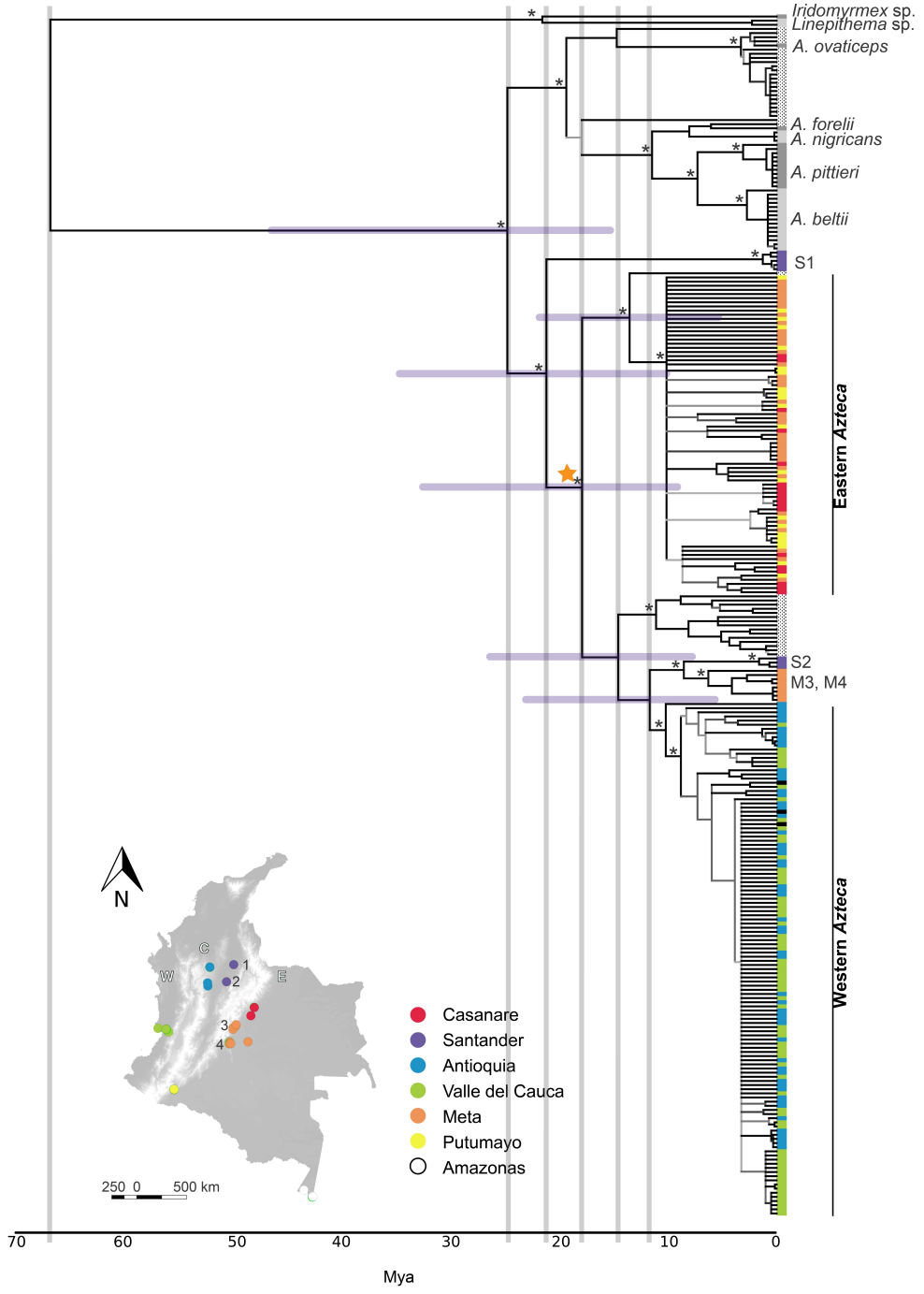


Figure 2.13: Calibrated majority consensus tree obtained from the large dataset of *Azteca* ITS2 sequences including sequences from NCBI (in grey), sequences from *Cecropia*-associated *Azteca* collected in Colombia (represented by dotted filling) and *Forelius* sp. and *Dorymyrmex* sp. outgroup sequences. Specimens in black are *Tococa*-associated *Azteca* not included on the species delimitation analyses. The star indicates the split between Western and Eastern *Azteca*. Support values correspond to the posterior probability of the branch and only values between 0.5 and 0.7 are shown. Black branches have a support higher than 0.7, branches with less than 0.4 posterior probability are collapsed and dotted branches have a support between 0.4 and 0.6. Asterisks indicate branches with a posterior probability higher than 0.9. The map shows the collecting sites and the color code for the specimens. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

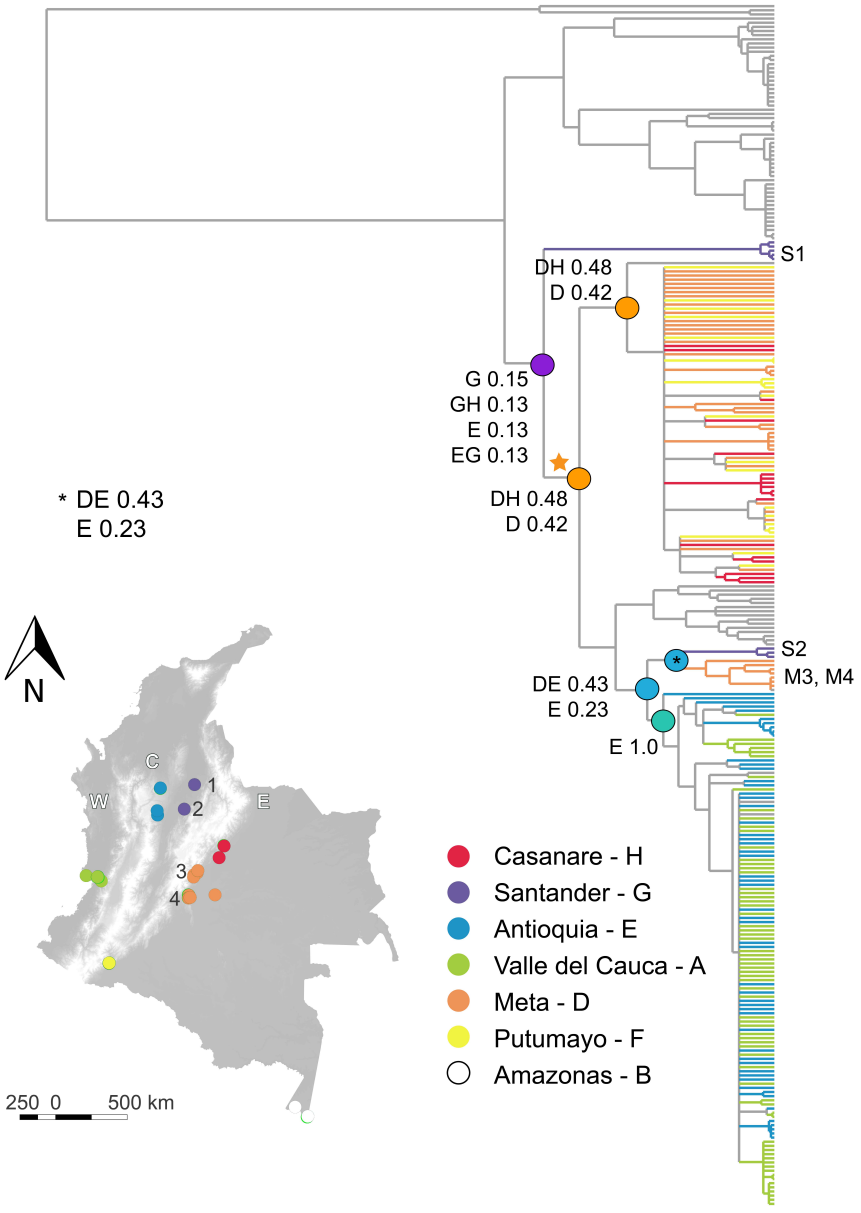


Figure 2.14: Calibrated majority consensus tree obtained from the large dataset of ITS2 sequences and the reconstruction of ancestral areas for the nodes of *Tococa*-associated *Azteca*. The two (or more) most probable ancestral areas and their probabilities are shown for the nodes of interest. The star indicates the split between Western and Eastern *Azteca*. The map shows the collecting sites and the color code for the specimens. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

2.6 Discussion

In this section, I will first discuss fieldwork observations of ants that inhabit *T. guianensis* across populations situated on opposite sides of the Andean cordillera. Then, I will discuss the identification of *T. guianensis* ants using DNA barcodes and species delimitation approaches under a phylogenetic framework. Finally, I will compare the timing of population divergence events with the timing of Andean uplift, and how such geological changes might relate to the geographical structure of the ant populations.

Sampling revealed that *Azteca* and *Pheidole* ants are the most common ant associates of *T. guianensis*. In most cases, both ant genera were found inhabiting the host in sympatry, although *Azteca* tended to be more common than *Pheidole*. The two exceptions are Valle del Cauca, where *Pheidole* was not collected, and Chocó, where *Azteca* was not found inhabiting *T. guianensis* (Table B.3 in Appendix B). Instead, *Azteca* was found inhabiting other Melastomataceae myrmecophytes (collected to test whether *Azteca* was not present in the area) and other plant families (*e.g. Cecropia*). Alvarez et al. (2001) found similar densities of *Tococa*-ant inhabitants in Chocó, collecting *Pheidole* in 90% of the surveyed *Tococa* and *Azteca* in only 2% of cases. Other studies of ant occupancy of Melastomataceae in South America (not including the Northern Andean region) report *Pheidole* as the most common inhabitant of *Tococa* (Fowler, 1993; Vasconcelos, 1991) and *Conostegia* (Alonso, 1998), but studies focusing in northern South America and Central America report *Azteca* as the main plant occupant (Cabrera and Jaffé, 1994; Bizerril and Vieira, 2002; Goitia and Jaffe, 2009). The sample collection protocol designed for this study focuses on *T. guianensis* inhabitants, and data does not provide evidence of the absence of ant species from an area or their ability to colonize other hosts. For this, it is necessary to survey all the ant species inhabiting plant hosts

and all the hosts present. *Crematogaster*, *Allomerus*, and other ant genera collected less frequently can be opportunistic ants inside *Tococa* domatia, but this can only be confirmed with ecological experiments and more observations.

2.6.1 DNA Barcoding

Delimitation of hypothetical *Azteca* MOTUs is possible using molecular markers, with the mitochondrial COI marker exhibiting congruent but more geographic structure than the nuclear ITS2 as expected from their different inheritance mechanisms. However, *Azteca* species representation in reference databases hindered the molecular identification of the specimens collected for this study. Molecular identification of specimens is useful when existing databases of voucher sequences contain all or most species. Otherwise, matches obtained to an incomplete set of reference sequences can only give an approximation to the taxonomy of the query. The resolution obtained from the COI sequences in my study was enough to identify specimens accurately to genera and to confirm morphological identification at this level. ITS2 sequences were equally successful at providing confirmation at the genus level, although successful identification at the species level was poor due to the low geographic coverage and species representation of ITS2 in the reference databases. For *Azteca*, ITS2 has the highest number of reference sequences, mostly from specimens collected in Costa Rica and Central America (Pringle et al., 2012). The genetic differences between these and the sequences from this study resulted in poor blast scores but also highlights the lack of sampling in areas like Colombia, a region that lacks voucher specimen-linked sequences of known Linnaean species. Moreover, my results highlight the importance of diversity surveys to increase sampling

of DNA, specimens and sequence collections. Hopefully, making available the sequences resulting from this study will contribute to the sampling of diversity in general.

The performance of the mitochondrial marker is better than the nuclear marker. As expected from a coding region, the COI alignment was unambiguous and easiest to sequence. However, COI results should be carefully interpreted as substitutions in the mitochondrial genome are fixed faster and the organelle's demographic patterns can be different from that of the nuclear genome. Unlike COI, ITS2 alignment was challenging, but it is more representative of species-level processes as opposed to population-level processes. ITS2 has been reported as useful for phylogenetic analyses at species level (Campbell et al., 1994; Gómez-Zurita et al., 2000; Alvarez and Hoy, 2002; Li et al., 2010). However, the presence of indels hinders the accuracy and homology of the segments alignment (Coleman and Vacquier, 2002), and has the potential to introduce noise in downstream analyses. The results of this study demonstrate that ITS2 provides enough resolution to cluster sequences into hypothetical MOTUs and that the broad geographic patterns emerging from ITS phylogenetic analyses are consistent with those for COI, despite the presence of indels in the ITS sequences. Moreover, these results prove the utility of ITS2 to identify diverged lineages in *Azteca*. ITS2 has enough resolution to identify two well-sampled MOTUs that correspond to Eastern and Western *Azteca*. The genetic distance between these lineages is considerable and both can be considered as two hypothetically different species. Within each of these MOTUs, the geographic structure in COI is a consequence of the smaller population size expected for mitochondrial DNA. Other MOTUs identified here have fewer specimens than Eastern and Western *Azteca*; however, they likely represent other rarely collected hypothetical species.

2.6.2 Phylogenetic inferences

Phylogenetic reconstructions were consistent for both small and large COI datasets and were generally concordant with results for the large ITS2 dataset. Two broad geographic patterns emerge from the phylogenies: *Azteca* sequences from Colombia form a group separate from the NCBI sequences (mostly from Central America), and the *Tococa*-associated *Azteca* form two distinct clades at the western and eastern flanks of the Andes. I observed few inconsistencies between the COI and ITS2 phylogenies, and between the small and large ITS2 phylogenies. The topological inconsistencies between COI and ITS2 can be attributed to the coalescent variation expected between any two gene trees within the same species tree or to mitochondrial capture. Similar incongruence between mitochondrial and nuclear markers have been reported for other *Azteca* sequence datasets (Pringle et al., 2012). In their study of *Azteca* inhabiting Central American *Cordia nodosa* plants, Pringle et al. (2012) found that mitochondrial and nuclear gene trees agreed in the allocation of individuals to species but were discordant in allocating individuals to populations within species, and phylogenetic relationships between individuals depended on the kind of data used. Regarding the small and large ITS2 datasets, it is also possible that the inconsistent placement of some branches in the small ITS2 dataset results from the reduced species sampling with respect to the large ITS2 dataset. Gene tree incongruence in *Azteca* species relationships highlights the coalescent variance between loci and is expected from groups that diverged recently. Multilocus coalescent approaches are more suitable for the analysis of discordant gene sets and will be applied to genomic-scale data in Chapter 3.

Overall, four groups of sequences can be recognized: the outgroups **(O)**, a group of only *Cecropia*-associated *Azteca* **(C)**, a group with other *Cecropia*-associated *Azteca*

sequences and NCBI sequences **(A)**, and a group of *Tococa*-associated *Azteca* **(T)**. The relative positions of **A**, **C** and **T** vary across COI and ITS2 phylogenies. COI recovers a polytomy of the four groups that informative variation in ITS2 resolves. Like observed by Ayala et al. (1996), *Cecropia*-associated *Azteca* do not form a monophyletic group according to COI sequences and as ITS2 sequences confirm. Within **T**, the Meta samples from M3 and M4, and the Santander samples S1 and S2 show conflicting topologies. The relative position of the M3 and M4 groups is inconsistent between the COI and ITS2 large datasets. Opposite to ITS2, COI recovers the position of these within the eastern clade, as also recovered by the ITS2 small dataset. Such incongruence can be the result of incomplete lineage sorting, recent admixture events, or a continued gene flow during population divergence. M3 and M4 are located in an area where the Andean Cordillera is lower than 2,000 m.a.s.l. in average (Figure 5.1 in Chapter 5). It is possible that gene flow between populations nearby those areas continued during and after the uplift as height was not a limitation, especially assuming continuous patches of forest. More recent admixture could explain the differences between trees; however, the precise inference of gene flow and population demography must ideally make use of more loci.

The position of S1 and S2 with respect to the rest of the samples suggests the presence of two distinct groups. The S1 is population is the sister group of all *Tococa*-associated *Azteca* while S2 is more closely related to the Western clade. The Central and Western cordilleras are lower towards the north and do not completely divide the area as the Eastern Cordillera does. Thus, gene flow between S1 and S2 Santander populations is possible, like how gene flow between Antioquia and Valle del Cauca can occur via the lowlands in the north of the Central and Western Cordillera. Because S1 is the sister group to all *T. guianensis*-associated *Azteca*, it is possible that *Azteca* populations on

each flank diverged from an ancestral species distributed over the Santander area as a result of vicariance. Both the COI and the ITS2 large datasets confidently recover their phylogenetic positions, and the conflicting position of S2 in the ITS2 small dataset tree can be attributed to similar processes as in the case of M3 and M4 as discussed later in this chapter.

Moving towards the tips and within the Western and Eastern clades recovered by COI and ITS2, COI recovers strong geographic structure in the populations resembling the collecting sites, a structure not detected in the ITS2 phylogenies. COI is expected to evolve fast enough to differ between species, but isolation by distance, a low effective population size (one quarter that of a nuclear locus) and the lack of recombination tend to enhance phylogeographic signal (Funk and Omland, 2003; Cognato, 2006; Meier et al., 2006; Schmidt and Sperling, 2008; Dupuis et al., 2012). The nuclear data, however, support the hypothesis that the *Azteca* entities are different across the Andes, but remain the same along the same flank.

Finally, the conflicting position of clade **C** of *Cecropia*-associated *Azteca* collected in Colombia is also likely to be caused by incomplete lineage sorting. All **C** samples were collected in Chocó and Tolima (the area between the Central and Western cordilleras), which is consistent with their position within the Western clade. Despite the position of clade **C**, *Azteca* collected from *T. guianensis* always forms a different clade from those collected from *Cecropia*, suggesting some host preference at the species level that can be further investigated. This pattern is clearer in the **A** and **C** clades with all COI sequences from *T. guianensis*-associated *Azteca* from NCBI and Colombia (with the only exception of one NCBI sequence of *A. quadraticeps*) clustering together. In addition to the predominantly *Cecropia* ants *A. ovaticeps* and *A. alfari*, other species

in clade **A** also associate with *Duroia* (Rubiaceae) and *Cordia* (Boraginaceae) (Ayala et al., 1996; Longino, 2007). The question of which factors might contribute to such host preference rises as these host plants have a similar distribution to *T. guianensis* throughout the Neotropics.

2.6.3 MOTU delimitation

All the methods for specimen delimitation used consistently group *T. guianensis*-associated *Azteca* specimens into seven monophyletic MOTUs. From those, two major MOTUs were consistently delimited using COI and ITS2 sequences, each MOTU corresponding to populations at opposite sides of the Eastern Cordillera (Western and Eastern *Azteca* MOTUs). Variation in number of MOTUs delimited by the methods corresponds to the inclusion or exclusion of singletons into other MOTUs and are a result of varying thresholds. Intraspecific distances within clades represented by singleton sequences cannot be estimated, blurring the limit between intra and interspecific genetic distances. Thus, membership of singletons cannot be resolved without increasing the sampling, a difficult task when the sampling does not include the diversity of the taxa of interest. On the other hand, substitutions accumulate at different rates across populations and markers, depending on their evolutionary history (Blaxter, 2004). Therefore, a delimitation threshold for nuclear markers is not necessarily the same as the threshold for mitochondrial markers. However, delimiting MOTUs using DNA barcodes proved to be useful to reveal the presence of geographically restricted *Azteca* MOTUs. Even though delimitation methods often require up-front decision making, they become useful when decisions are informed by independent sources of information such as geography,

morphology or ecology (Smith et al., 2005; Jones et al., 2011; Puillandre et al., 2012; Ronque et al., 2016; Sukumaran and Knowles, 2017).

The Western and Eastern *Azteca* MOTUs include most specimens and are congruent with all phylogenetic reconstructions supporting the hypothesis of two possible *Azteca* species, one on each side of the Andes. The 1.66% JC69 threshold value reported by **ABGD** for COI and the cut-off value of 7.13% sequence divergence reported by **jMOTU** are consistent with reported threshold sequence divergence values of 1-3% and up to 15% for COI (Hebert et al., 2003b, 2004b; Smith and Fisher, 2009; Smith et al., 2013; Bribiesca-Contreras et al., 2013; Stahlhut et al., 2013). Lower cut-off values resulting in smaller COI MOTUs reflect the strong geographic structure of the populations and might overestimate the number of final MOTUs. **ABDG** is particularly sensitive to population structure as the method estimates threshold values as the cut-off at which no further division of MOTUs can take place.

The lack of support for the delimited MOTUs and the high posterior probabilities reported by **BPP** for the *a priori* grouping of samples is likely due to topological inconsistencies between COI and ITS2 and the marked geographic structure of the COI data. In a recent study, Sukumaran and Knowles (2017) discuss species definition under the multispecies coalescent model as implemented in **BPP**. Based on simulated data they conclude that **BPP** performs better at detecting structure in the data than at delimiting species from that structure, which will cause **BPP** to overestimate the number of species. The application of a threshold to define species does not necessarily reflect the process of speciation; species do not appear instantaneously from structured populations and the process often involved more than just genetic differentiation. Speciation requires genetic differences to remain and accumulate in time, typically involves

reproductive isolation via biological barriers or geographic isolation, and can involve niche partitioning when in sympatry. However, it is an approach that makes it possible to define discrete homologous units, especially when applied to closely related lineages undergoing similar evolutionary processes. This simplifies analyses and interpretation in studies of lineage diversification. Further studies assessing reproductive isolation are needed before establishing the homology of these MOTUs to species.

2.6.4 The timing and geographic pattern of *Tococa*-associated *Azteca* divergence

Fossil calibrations suggest that the split between *Azteca* and the outgroup genera occurred somewhere between 25 and 67 Mya and that the tMRCA to crown *Azteca* dates from around 26.33 Mya (±3.57, averaging across the models tested). Similarly, Ward et al. (2010) estimates the divergence between *Azteca*+sister genera and *Dorymyrmex* to 50 Mya (95% CI= 41-60 Mya) and Moreau et al. (2006) estimates the same divergence between 65-75 Mya. The estimated tMRCA of all *Azteca* included here is older than reported by other studies: tMRCA to all *Azteca* is 12 or 14 Mya (95%CI=7-22 Mya), depending on the method used by Ward et al. (2010); tMRCA to all *Cordia*-associated *Azteca* between 10-22 Mya, according to Pringle et al. (2012). Nevertheless, the means (and medians) estimated across all models fall within the confidence intervals those studies report. Compared to my study, Moreau et al. (2006) and Ward et al. (2010) use more loci and fossils, but their analyses are based on concatenated alignments and the total sampling of *Azteca* corresponds to three sequences. *Azteca* is a genus with as many as 100 species and subspecies, and in all the above examples, the phylogenetic reconstructions are poorly sampled.

The divergence between the Western and Eastern *Azteca* MOTUs is estimated to be around 17 Mya (± 4.27), not long after the origin of *Azteca*. This suggests that the split occurred well before the Andes reached its maximum height but after the three cordilleras were formed and their height was 40% of their current altitude. Evidence suggests that by 11.8 Mya the Eastern Cordillera was already a continuous range (Hoorn et al., 1995), and it is likely that the closure of the Eastern Cordillera contributed to the western-eastern split in *Azteca*. The genus is restricted to lowland forest areas and it is possible that the height of the Andes 12 Mya, especially the Eastern Cordillera, was enough to restrict their altitude range and act as gene flow barrier throughout most of the cordillera's length. The data suggests that such gene flow could have continued only in areas where the Andes was lower than 2,000 m.a.s.l. and where continuous corridors of forest could allow the dispersal of the ants.

The results of this study show strong evidence supporting the hypothesis that the Andes, especially the Eastern Cordillera, played a major role in population divergence by promoting vicariance and by limiting gene flow. This pattern seems to be consistent with other insect taxa distributed around the Andes. A study on arboreal ants of the genera *Camponotus*, *Dolichoderus* and *Ectatomma* collected in both sides of the Ecuadorian Andes (in the Ecuadorian Chocó and Ecuadorian Amazonas) showed splits within species forming a clade from the Chocó and another clade from Amazonas. A calibrated phylogeny revealed the split to have occurred between 20-5 Mya, coincident with the period of highest and fastest activity of mountain uplift (Troya, 2012). In the case of *Azteca*, it is the highest Eastern Cordillera that has more effect on the evolutionary history of the genus, the Central and Western Cordilleras have lower altitudes and do not disconnect the entire northern territory, allowing for some gene flow going through the northern lowlands.

The outgroup placing of S1 and Central American sequences support the hypothesis of *Azteca* migrating from the north-west to the northern Andes. Fossil and biological evidence support an early emergence of the Panama isthmus, starting in the Oligocene to Miocene transition (around 23 Mya) and extended until around 10 Mya (Bacon et al., 2015). Thus, conditions were set for *Azteca* to migrate north to south, however, that hypothesis is not tested in this study. Moreover, uncertainty in the ITS2 phylogeny is likely to introduce noise in the estimation of ancestral areas (Figure 2.14), and the estimation of Santander and Antioquia as ancestral areas to all *T. guianensis*-associated *Azteca* can be an artefact of the outgroup placing of S1 and other Central American NCBI sequences.

To summarize, the results from this chapter revealed two major separate *Azteca* lineages associated to *T. guianensis* which likely diverged due to the Andean uplift and a reduced gene flow between populations. Both geographic patterns of distribution and the congruent timing between the uplift and the divergence events are strong evidence of the Andes promoting diversification within *Azteca*. Few nuclear and mitochondrial incongruences were found, possibly due to continued gene flow in areas where the altitude is not high enough to limit gene flow between ant populations. But whether this is a pattern of continued gene flow or a consequence of mitochondrial capture will be explored in Chapter 3 using genomic data from the populations revealed in this chapter.

CHAPTER

3

GENE TREES, SPECIES TREES AND PHYLOGENOMICS OF *AZTECA*

3.1 Introduction

Studies comparing phylogenetic reconstruction methods show how multi-locus approaches perform better than methods using single or very few loci (Pamilo and Nei, 1988; Brito and Edwards, 2009; Degnan and Rosenberg, 2009; Yang and Rannala, 2010). Even combining one nuclear and one organelle markers might not be enough. Because nDNA and organelle DNA are inherited differently, have different effective population sizes and demographic histories, using one loci from each does not necessarily tell the complete

story of its genome. Chapter 2 shows that COI and ITS2 tell conflicting stories regarding two ant populations from Santander (referred there as S1 and S2), and both markers are not enough to resolve their phylogenetic positions. Thus, using the advantages of whole genome data, I assess possible causes of tree discordances between nuclear and mitochondrial DNA and address the relationships between S1 and S2 respect to *Azteca*. Observing the patterns of topological incongruence between both genomes can reveal events of *e.g.* mitochondrial capture or hybridization. Using these next-generation data, I also estimate more accurately the age of Western and Eastern *Azteca*.

3.1.1 Species tree and gene trees

Speciation results from a continuous process of population isolation through time by geographic distance, disruptive selection, genetic drift, or other mechanisms that increase the genetic distance between populations. Thus, genetic divergence among such populations is a pre-requisite for either allopatric or sympatric speciation to take place. It is possible to examine such genetic divergence and subsequent levels of speciation using phylogenetic or coalescent methods. Traditionally, phylogenetic methods estimate gene trees from one or a few loci that are assumed represent the species tree; however, few gene trees provide a partial and sometimes misleading reconstruction of the species evolutionary history (Doyle, 1992; Page and Charleston, 1997; Degnan and Rosenberg, 2009).

The estimation of the origin of polar bears is an example of misled conclusions derived from the use of few loci. Using only mitochondrial genomic data, Lindqvist et al. (2010) estimated the time of branching of the polar bear to be around 100-166 thousand years ago. Both Lindqvist et al. (2010) and Davison et al. (2011) found extant polar bears

nested within brown bears and as a sister lineage to the Alaskan brown bears. These results even agreed with previously published mtDNA phylogenies incorporating 14 nuclear loci (Pagès et al., 2008). It was not until Hailer et al. (2012) included longer, multiple nuclear loci from across the genome that the previous assumptions about the polar bear's origin were more accurately estimated. Hailer et al. (2012) found that polar bears are a distinct, older lineage than brown bears. Later, Cahill et al. (2013) confirmed the polar bears as sister lineages to brown bears and revealed admixture after divergence and gene flow predominantly from polar to brown bears using full genomes. The polar bear case illustrates the power that genomic data and more comprehensive analyses provide reliable stories about the species evolution, rather than limiting the stories to that of few genes.

Costs of sequencing, time, and a restricted knowledge of potentially informative regions of the genome were limiting factors in phylogenetic analyses. But recent developments in next-generation sequencing technologies and the decrease in costs allow for the sequencing of thousands of loci, enabling more accurate species tree estimations. However, having thousands of loci can be disadvantageous as it potentially increases the complexity of the multiple evolutionary trajectories whilst trying to make comprehensible assessments of them. Such complexity is visible when topology conflicts occur among genes and between genes and species tree topologies. Among the causes of such conflicts are Incomplete Lineage Sorting (ILS), hybridization, mito-nuclear discordance, gene flow, recombination and gene duplication, some of these I will be discussed later in the chapter.

3.1.2 Estimation and calibration of species trees

To understand why conflict between genes and species trees occur, it is necessary to understand how species trees are estimated. Concatenation and coalescence are widely used methods to reconstruct species tree from gene trees (Liu et al., 2015b). Concatenation methods use a super-matrix of all genes, treating them as a single unit assuming a shared evolutionary history and no recombination. Because this method estimates parameters for one matrix, it is faster than coalescence. It also performs relatively well if the internal branches of the tree species are long and there is little or no conflict between gene trees. However, concatenation does not consider the varying histories of different genes. For instance, if genes with different evolutionary and recombination rates are concatenated in one matrix, the assumption of a homogeneous gene history for the concatenated matrix is violated. Moreover, if only linked genes are included, the resulting species tree represents the tree of the linked genes but misses that of the remaining of genes throughout the genome. Thus, concatenation can bias the species tree and result in the overestimation of branch supports as fewer genes (and therefore less variance) are introduced during the species tree estimation (Kubatko and Degnan, 2007; Liu et al., 2015b,a). Conversely, coalescence methods treat genes independently and integrates their different histories into the species tree estimations. A commonly used coalescent method is the multispecies coalescent model, which estimates divergence times and events between multiple species (Rannala and Yang, 2003). Because this method uses a matrix per gene (concordant or conflicting) instead of one concatenated matrix, it is robust to high levels of ILS (Kutschera et al., 2014; Angelis and Dos Reis, 2015; Davidson et al., 2015). Applications to this method first estimate all

gene tree topologies, then the species tree is obtained by summarizing gene tree statistics or by Bayesian or likelihood methods. Software such as **ASTRAL** (Mirarab et al., 2014) and **STEM** (Kubatko et al., 2009) use summary statistics to consistently recover the true species tree even with topology conflicts, performs well when the number of loci is high and does not exceeds computational requirements. Other software like **BEAST** (Drummond et al., 2012) and ***BEAST** (Heled and Drummond, 2010), perform better but are computationally very expensive for large numbers of loci (Zimmermann et al., 2014; Liu et al., 2015a).

3.1.3 Species and gene tree conflict

Different populations undergo different demographic processes that affect how, and which gene variants are passed through generations. Thus, differing histories among genes throughout the genome will cause gene trees to be discordant and potentially different from the species tree. Nonetheless, those differences can result in similar patterns of gene discordance that may only be distinguishable by hypotheses testing. In cases where the demographic scenarios produce very similar patterns, *e.g.* hybridization and retention of ancestral polymorphisms, distinguishing between both cases requires multiple loci to provide sufficient statistical power.

Tree calibration and gene-species tree conflicts

Other than topological, temporal tree conflicts arise as absolute diverging times are different among genes. Lineage divergence is a continuous process resulting from sequential gene divergence through time to the point at which lineages become genetically isolated. Estimating species divergence times using one or a few loci can bias node age

estimates. For instance, copies of a gene in two diverged lineages are likely to coalesce in time t before the lineage divergence time T when looking forward in time. Depending on the ancestral population size at T, the t of a single gene can be a valid approximation to the time of lineage divergence so that if T is large and the ancestral population is small, t will be close to T as coalescence occurs faster in smaller populations. Conversely, when T is short and the ancestral population size is large, gene coalescence is slower and T and t will be very different. An additional effect of a short T and large ancestral population sizes is the increased probability of gene-to-species tree conflicts, which in turn increases the difference between T and t producing age estimates that will be less likely to be correct when using only one or two loci. Thus, depending on the evolutionary history of that gene its tree may or may not match the diverging times of the species tree.

Species and gene tree conflict: incomplete lineage sorting

ILS occurs when one gene copy from one population coalesces with another copy that is not from the same population or its closest relative, but a less closely related one: it is the failure of lineages in a population to coalesce (Degnan and Rosenberg, 2009). It can result from the segregation of an ancestral polymorphism causing a mismatch between species and gene trees (Pamilo and Nei, 1988). Furthermore, probabilities of ILS increase when the time between species divergence events is short and population sizes are large, as the expected time to coalescence events (t) increases (Liu et al., 2015a). ILS can be distinguished from other causes as the coalescence of the incongruent gene is assumed to occur before the splitting time of the species (looking forwards in time), however old coalescence events are not unique to ILS and old hybridization events

can produce the same patterns (Joly et al., 2009). Similar patterns to ILS are also caused by gene duplication and recombination. Gene duplications contribute to gene tree conflict as both copies can experience completely different evolutionary trajectories and copy selection for tree reconstruction is uninformed. Thus, it is important to identify orthologue and paralogue genes across genomes to reduce potential causes of ILS in phylogenetic reconstructions. Ancient recombination events between closely related lineages are more difficult to distinguish from ILS and require a more detailed study of the entire genome and other genes showing the same pattern.

Species and gene tree conflict: hybridization

Hybridization is common in many taxa and it plays an important role in generating biotic diversity and promoting rapid adaptation (Grant and Grant, 1996; Ellstrand and Schierenbeck, 2000; Rieseberg et al., 2003; Arnold, 2004). Hybridization occurs when closely related species that have not acquired reproductive barriers come into secondary contact or are sympatric (Mallet, 2009). For instance, species of *Pogonomyrmex* and *Solenopsis* ants produce hybrid workers among co-generates (Meer et al. 1985 and more references in Feldhaar et al. 2008). Incongruence patterns are potential indicators of recent hybridization events when the incongruent genes coalesce after the species divergence, but these patterns can be confounded with ILS or gene duplications if the hybridization event is older than the species divergence (Linder and Rieseberg, 2004). Moreover, migration of individuals between isolated populations and secondary contact produce similar incongruent patterns as hybridization.

Species and gene tree conflict: discordance between mitochondrial and nuclear genomes

Mito-nuclear discordance refers to incongruent nuclear and mitochondrial topologies that results from mitochondrial capture, the idiosyncrasy of some mating systems or non-random mating (Eyer et al., 2016), sex-biased dispersal (Roca et al., 2005; Petit and Excoffier, 2009), genetic drift when dispersal is low (Bonnet et al., 2017), and selective sweeps such as those associated with some bacterial symbionts such as *Wolbachia* (Toews and Brelsford, 2012; Ilinsky, 2013). Mito-nuclear incongruence is expected as mitochondria is inherited uniparentally and have smaller effective population sizes than nDNA (Charlesworth, 2009). Moreover, it can arise in the presence or absence of geographic isolation, via secondary contact following isolation or via selection (Irwin, 2002; Toews and Brelsford, 2012).

Mito-nuclear discordance has been observed in ants before. For instance, in genetically different lineages of *Cataglyphis hispanica* dessert ants that coexist sympatrically. Their colonies include clone workers product of asexual reproduction whose genetic profile matches that of the queen ant. But the colony also sexually produces workers that are hybrids between the coexisting lineages, causing mismatching between the nuclear and mitochondrial phylogenies (Eyer et al., 2016). Evidence of mito-nuclear discordance was found in *Azteca* when comparing the relative position of the Santander *Azteca* (S1 and S2) obtained from COI and ITS2 (Figure 2.7 in Chapter 2). When looking at the small datasets of only *Tococa*-associated *Azteca*, the nuclear ITS2 suggests that the S2 population is closer to Eastern than to Western *Azteca*, whilst the mitochondrial COI suggest that S2 is closer to Western *Azteca*.

As discussed above, the accuracy of species tree estimations and time calibrations increase with the number of loci used, which in turn allows for the sampling of the subtle differences in alternative population histories and demographic parameters. In Chapter 2, I identified the *Azteca* ants associated with *Tococa* and demonstrated a west-east structure between populations with respect to the Eastern Andean Cordillera. In that chapter I identified incongruences between mitochondrial and nuclear markers in a population of Santander. The aims of this chapter are *(1.)* to produce low coverage ant genomes from the populations identified in Chapter 2, including two specimens from Santander; *(2.)* to elucidate the mito-nuclear incongruence patterns or determine if incongruence also occurs among nuclear loci; *(3.)* to estimate the tMRCA to all *T. guianensis*-associated *Azteca* lineages using genomic data derived from the assemblies; and *(4.)* to explore the presence/absence of the bacteria symbiont *Wolbachia* and reconstruct its species tree, patterns of divergence and its congruence respect to the *Azteca* host. This last aim takes advantage of the identification of contaminant reads in the *Azteca* genomic data.

Thus, this chapter provides the first low coverage assemblages of *Azteca* useful in more in-depth demographic analyses beyond those presented in this chapter. Moreover, exploring the congruence among genes in *Azteca* lineages provides a better understanding of their divergence process and facilitates hypotheses formulations. These analyses can result in two possible scenarios. In the first one, incongruence occurs between mitochondrial-like and nuclear loci but does not occur within mitochondrial and nuclear groups. This pattern is an indicator of either mitochondrial capture or hybridization between Eastern *Azteca* and Santander1 populations. In the second scenario, incongruence occurs among mitochondrial-like and nuclear loci, but also within each genome group (especially among nuclear loci). This is an indication of ILS or other mechanisms

but mitochondrial capture. Mitochondrial and nuclear loci differ in their inheritance, ploidy, effective population size and evolutionary rates; looking at the evolutionary history of each genome provides insight in favor of unidirectional gene flow, possible hybridization, plastid capture and other events during evolution. Additionally, when mito-nuclear discordance is suspected, separate analysis of mtDNA and nDNA reduces the computational resources required for the analyses to converge, which can be a problem when using genomic data. I expect to understand possible causes of the patterns observed in Chapter 2 to be able to formulate appropriate hypothesis in future analyses. Estimating the tMRCA of all *Azteca* lineages from genomic data will confirm the results from Chapter 2, which only uses ITS2 and COI. Finally, if *Wolbachia* reads can be recovered from the *Azteca* assemblies, the patterns of diversification of this vertically transmitted symbiont can serve as a point of comparison for those patterns found between the horizontally transmitted *T. guianensis-Azteca* mutualism. This alpha-Proteobacteria is best known for its associations with Arthropods and Nematodes, and for influencing how its host reproduce (Werren et al., 1995; Werren, 1997).

3.2 Methods

3.2.1 DNA extraction and library preparation

Based on the results provided in Chapter 2, I selected fifteen *Azteca* individuals that showed no evidence of ambiguous MOTU membership from three populations to the west and four populations to the east of the Eastern Cordillera. DNA was extracted from soldier ants following the Qiagen DNeasy Blood and Tissue kit protocol with modifications to increase the yield and purity needed for further library preparation.

The whole ant body was homogenized and 180μL ATL buffer and 20μL Proteinase K (10μg/ml) immediately after. The mix was left incubating overnight at 37°C. Then, 1μL RiboShredder™RNase Blend (Illumina) was added to the mix and left incubating for 30 min at 37°C. The protocol was then followed according to the manufacturer´s instructions. Highly degraded DNA fragments were removed at the DNA elution step by adding 50μL of EB buffer to the column followed by centrifugation. To obtain as much DNA as possible, DNA was eluted from the column by applying 40μL EB buffer, then the flow-through was loaded back to the column, incubated and centrifuged. A final elution was done with 25μL EB buffer. DNA integrity was assessed on a 2% agarose gel stained with 2.5μL ethidium bromide and quantified using the high sensitivity assay of QuBit DNA quantification system (Invitrogen).

DNA was fragmented using a Bioruptor Plus (Diagenode, Belgium) for 5 low power cycles of 25 sec on/90 sec off. Samples with higher molecular weight were sonicated for 5 extra cycles of 20 sec on/90 sec off. Resulting fragment sizes and weight was measured using the High-sensitivity D1000 protocol (Agilent 2200 TapeStation system, United States). Library preparation followed the TruSeq Nano LT protocol for 350bp insert size using 100ng input DNA per sample (Illumina, FC-121-4001). During the last clean up step for amplified DNA, 37.5μL Sample Purification Beads were used instead of the 50μL indicated in the protocol. Library quality was assessed on the TapeStation and final sample concentration measured using the high sensitivity assay of the QuBit DNA quantification system. Equimolar amounts of all samples were pooled together and sequenced on one lane of the Illumina HiSeq 4000 platform. Whole genome sequencing of the three Amazonian individuals was carried out as described above, but using 200ng DNA from *Azteca* ants and following the TruSeq Nano LT kit for DNA samples (Illumina) for 550bp insert sizes. These samples were sequenced following a

different library preparation protocol as they were the first genomes to be generated and were meant to help as reference genomes for the rest of the samples if needed. Library quality was assessed on an Agilent 2100 Bioanalyzer system (Agilent Technologies, United States). Samples were pooled and sequenced on one lane of the Illumina HiSeq 2500 platform in high-output mode. Both sequencing runs were carried out by the Edinburgh Genomics facility.

3.2.2 *Azteca de novo* genome assembly

Genome assembly followed similar steps as those used to assemble the *Tococa* genomes in Chapter 4. Reads were trimmed, and Illumina adapters removed with **Trimmomatic** (Bolger et al., 2014). Quality of the reads before and after trimming was assessed with **FastQC** (Andrews and others, 2010). A preliminary **Velvet** (Zerbino and Birney, 2008) assembly, information about read coverage obtained from mapping the reads back to this assembly (using **BWA-MEM v0.7.15**, Li and Durbin 2009), and taxonomic information from **Blast** hits querying the contigs were used to assess contamination with **Blobtools v0.9.19.5** (Kumar et al., 2013; Dominik R. Laetsch, 2017). Contaminant reads were removed and the remaining reads assembled using **MetaSPAdes v3.10.1** (Nurk et al., 2017). Assembly completeness and orthologue prediction were carried out using the Benchmarking Universal Single-Copy Orthologs (**BUSCO v2**, Simão et al. 2015) comparing the assemblies against the Hymenopteran orthologue database and using the honey bee as the default species parameters for **Augustus**. **BUSCO** identifies orthologous single-copy genes using **tBLASTm**, then predicts the gene structures using **Augustus** to finally assess the completeness of the predicted genes and classify the matches into "complete", "duplicated", and "fragmented", using HMMER hidden

Markov models. The Hymenoptera database has a total of 4415 **BUSCO** groups from 25 Hymenopteran species to use as template during the prediction step (Simão et al., 2015).

3.2.3 Maximum likelihood species tree

Complete single-copy genes shared across all 15 *Azteca* assemblies were aligned using **MUSCLE v3.8.31** (Edgar, 2004). Maximum-Likelihood estimations of the trees and branch lengths with 500 bootstrap replicates were conducted using **RAxML v8.2.9** (Stamatakis, 2014). Best bipartitions of each gene tree were summarized and branch support conducted using **ASTRAL v4.10.12** (Mirarab et al., 2014) to produce a species tree from all gene trees under a multi-species coalescent model. **ASTRAL** is more robust than other methods like Maximum Pseudo-likelihood for Estimating Species Trees (**MP-EST**, Liu et al. 2010) to differences in the input tree estimation method and to medium to high levels of incomplete lineage sorting in the data (Mirarab et al., 2014; Meiklejohn et al., 2016). The software reports a normalized quartet score (from zero to one) representing the concordance between gene trees and the estimated species tree (the higher the score the more congruent are genes and species trees; a lower score indicates *e.g.* incomplete lineage sorting). The branch scores are local posterior probabilities as a function of the number of genes, the frequency of different quartets of that branch (unrooted trees of four tips) and the minimal informative element in the tree (Zhaxybayeva et al., 2006).Branch scores can be interpreted as the frequency of a branch given the size of gene sampling: the more genes and the higher the branch frequency, the better the score is.

A **BUSCO** search for the *Linepithema humile* ant genome (NCBI Accession number ADOQ 00000000) and a Maximum Likelihood reconstruction of the gene trees using *L. humile* as outgroup were performed to generate a *Azteca* rooted species tree with **ASTRAL v4.10.12** (Mirarab et al., 2014). The genome of *L. humile* is the only Dolichoderinae ant assembly available and therefore the closest to *Azteca* (Wild, 2009). The use of an outgroup allows for the identification of rooted clades within the ingroup and therefore the concordance of those clades throughout the gene trees (Smith et al., 2015). Because high branch support values can mask significant conflict in multi-locus analyses (Salichos et al., 2014; Kobert et al., 2016), identifying genes that give conflicting topologies is important for understanding potential errors and biological processes (Walker et al., 2017). Using the **Phyparts** program (Smith et al., 2015) all unique bipartitions in a set containing all gene trees were compared against the rooted species tree to identify how many genes were producing conflicting topologies. Such topological incongruities can be a sign of systematic errors introduced during data processing, but also of hybridization, incomplete lineage sorting, and horizontal gene transfer (Bleidorn, 2017).

3.2.4 Identification of mitochondria-like loci

Mitochondria-like genes were identified with a **Blastx** search implemented in **BLAST v2.6.0** (Camacho et al., 2009), querying the single-copy genes identified with **BUSCO** against the curated protein database, then filtering the annotations of the hit with the best score. **BUSCO** identification of single-copy genes is based on conserved orthologous genes and I expect the mitochondria-like genes to come from the mitochondrion under the assumption that mitochondrial introgressions in the nuclear genome have

accumulated enough substitutions under lack of selection pressure such that **BUSCO** does not identify them as conserved genes. However, more detailed analyses are required to distinguish between true mitochondrial genes and mitochondrial genes introgressed into the nuclear genome that has not accumulated enough substitutions. For this reason, mitochondrial genes used here will be referred to as mitochondria-like genes. In the future, *Azteca* mitochondrial genomes will be assembled using either **MITObim** or **Norgal** (Hahn et al., 2013; Al-Nakeeb et al., 2017)

3.2.5 Gene tree node age calibrations

Divergence times between Eastern and Western *Azteca* across all genes were estimated and calibrated under the coalescent model using **BEAST v1.8.4** (Drummond et al., 2012), keeping nuclear and mitochondrial-like genes separated. Two important considerations about node age calibration are worth mentioning. First, fossil calibrations are more reliable than those based on estimated evolutionary rates (if the age of the fossil is certain). A well-identified *Azteca* fossil from the Dominican amber is usually used, but in the absence of other *Azteca* species there is not an alternative node to set the fossil calibration other than the node of interest. Setting the calibration in that node will bias the analyses. Second, the use of evolutionary rates relies on the certainty with which such rates are estimated. In addition, evolutionary rates can be different amongst lineages and genes (Richardson et al., 2001; Baer et al., 2007; Smith and Donoghue, 2008). Thus, a reported evolutionary rate for insects and a node time calibration derived from a Dolichoderinae fossil calibration were used in independent tests. For the rate calibration, a normally distributed hyperprior with a mean of 1.77 $\pm$0.19% per lineage per My for mitochondria-like genes and a mean of 1.84 $\pm$1.52% per My for

nuclear genes was set for all genes (Ho and Lo, 2013; Papadopoulou et al., 2010). For the fossil calibration, *Linepithema* was used as an outgroup and a normally distributed prior with a mean of 47 ±7 My was set for the split between *Linepithema* and *Azteca*, based on the fossil-calibrated phylogeny of Dolichoderinae (Ward et al., 2010). All analyses were carried out with two groups of 50 randomly selected single-copy genes from each of the mtDNA and nDNA datasets, using two independent chains of 500 million states for each test. Analyses were run using **BEAST** instead of ***BEAST**. The latest is designed to co-estimate species trees from gene trees in a coalescent framework but at least two individuals for each species must be included in the analysis (Heled and Drummond, 2008, 2010). Because there is not enough evidence suggesting S1 and S2 belong to the same species (as discussed in Chapter 2), S1 and S2 were treated as independent species and thus, ***BEAST** assumptions are violated. Finally, to evaluate the support of the different positioning of the S2 *Azteca*, the branch posterior probabilities of all **BEAST** gene tree datasets were plotted using **DiscoVista** (Sayyari et al., 2017). The clades evaluated are Western *Azteca*, Eastern *Azteca*, Santander *Azteca* (this is S1 and S2 populations within the same clade), Western+S2, and Eastern+S2.

3.2.6 *Wolbachia de novo* genome assembly

Reads with a significant hit to *Wolbachia* (Rickettsiales) are among the most frequent contaminants. To assemble the genomes of the *Wolbachia* strains present in *Azteca*, all contigs with a **Blast** hit to *Wolbachia* were filtered out from the contaminants and their taxonomic scores assigned by **Blobtools** manually inspected (a taxonomic score is assigned to the contig depending on the taxonomic **Blast** hit and how many taxonomic ranks are assigned by **Blast** to the contig). Only contigs with a single

taxonomic hit to **Wolbachia** or, in the case of multiple taxonomic hits, with the highest score for the hit to **Wolbachia**, were used. All sequenced reads were mapped to the **Wolbachia** contigs using **BWA-MEM v0.7.15** (Li and Durbin, 2009) and only those that mapped were used to assemble the genome using **MetaSPAdes v3.10.1** (Nurk et al., 2017). This process was done for each *Azteca* sample independently. A **BUSCO** search for single-copy genes was run for the *Azteca-Wolbachia* and all *Wolbachia* assembled genomes available in NCBI, using the Proteobacteria database. Genes were aligned with **MUSCLE v3.8.31** (Edgar, 2004), Maximum Likelihood trees estimated with **RAxML v8.2.9** (Stamatakis, 2014) and a species tree estimated with **ASTRAL v4.10.12** (Mirarab et al., 2014).

3.3 Results

3.3.1 *De novo* genome assembly

Whole genome sequencing of fifteen *Azteca* specimens resulted in around 50 million reads after removing low quality and contaminant reads, which accounted for up to 20% (Table 3.1). Kmer counts of all *Azteca* reads showed a normal distribution of kmer frequencies and no evidence of highly repetitive kmers (Figure C.1 in Appendix C). The N50 of *Azteca* **MegaSPAdes** assemblies is on average higher than 4 Kb with one exception (Table 3.2) and the **BUSCO** completeness is around 80% in all assemblies except for three, for which the **BUSCO** completeness is around 60% (Figure 3.1). In comparison, **BUSCO** detected 97.1% complete single-copy gene orthologues in the outgroup *Linepithema* genome. The assemblies with the lowest completeness percentage

are those with the highest N50, possibly due to incorrect merging of reads with repetitive segments that resulted in elongated contigs and misplaced gene fragments. Less than 1% of the genes in *Azteca* and 0.3% in *Linepithema* are duplicated according to **BUSCO**. However, this is not enough evidence for a lack of duplications in other genes across the genomes. After a **Blastx** search of the **BUSCO** single-copy genes against hymenopteran protein sequences in NCBI, a total of 125 genes showed best hits to a mitochondrial protein in Hymenoptera.

From a total of 57,043 contaminant contigs removed, half were identified as the alpha-Proteobacterium *Wolbachia*. Reads from those contigs were filtered out from the corresponding host's assemblies and a total of five *Wolbachia* genomes were assembled successfully, all of them from Western *Azteca* (Table C.1 in Appendix C). A total of 41 single-copy **BUSCO** genes (from a gene set of 221 **BUSCO** genes for Proteobacteria) are shared across the five *Wolbachia* and the reference *Wolbachia* genomes, and the **ASTRAL** tree from those genes group together the *Wolbachia* from *Azteca* in a clade within the **A** *Wolbachia* and sister to a clade of *Wolbachia* inhabiting *Drosophila* and *Nomada* bees (Figure 3.7)

Table 3.1: Number of Illumina pair end reads (in the order of million reads) before and after trimming and contaminant read removal.

	Total reads	**After trimming**	**After filtering**	**Coverage***	**%GC content**
MFT146	77.33	75.77	72.82	29.13	34
MFT151	73.14	70.05	62.11	24.84	33
MFT162	69.82	66.96	59.05	23.62	33
MFT327	68.12	65.07	63.51	25.40	35
MFT334	59.54	56.95	55.18	22.07	35
MFT400	58.50	56.08	45.99	18.40	34
MFT404	71.84	69.24	68.23	27.29	35
MFT472	66.60	66.60	52.60	21.04	34
MFT493-S2	65.71	63.39	52.67	21.07	34
MFT516	57.62	57.59	52.22	20.89	35
MFT517	66.43	66.39	58.66	23.47	35
MFT527	59.83	59.78	53.74	21.50	35

Table 3.1 continued from previous page

MFT538	59.02	58.94	51.56	20.62	36
MFT591	63.85	63.75	56.96	22.78	35
MFT602-S1	55.19	33.44	48.10	19.24	35

*Approximate coverage assuming all reads are 120 bases long (the distribution ranges from 20-150 bases)

Table 3.2: Statistics for the *Azteca de novo* genome assemblies produced by **MetaSPAdes** using filtered reads.

	Scaffold*	**No. of**	**Longest**	**Scaffolds**		**Contigs**	
	(Mb)	**scaffolds**	**scaffold (Kb)**	**N50**	**L50 (Kb)**	**N50**	**L50 (Kb)**
MFT146	294.08	107965	234.57	4174	16.19	26264	2.63
MFT151	288.95	112426	230.57	4409	15.12	26949	2.55
MFT162	285.88	118378	367.16	4476	14.62	30064	2.24
MFT327	276.61	121894	448.92	5406	11.2	22670	2.99
MFT334	278.49	130144	231.2	5440	11.14	21624	3.13
MFT400	271.22	131286	219.12	5584	11.07	24020	2.8
MFT404	280.58	153811	252.78	6101	10.24	29268	2.37
MFT472	274.83	132085	251.74	5889	10.61	26889	2.53
MFT493-S2	277.49	128275	198.6	6136	10.2	28199	2.45
MFT516	272.91	179671	271.42	7776	7.84	29885	2.26
MFT517	271.11	181864	205.79	8520	6.96	32259	2.08

Table 3.2 continued from previous page

MFT527	273.23	175939	271.81	733	8.08	28877	2.33
MFT538	267.76	304855	116.89	24054	2.69	34412	1.92
MFT591	269.39	147935	376.48	6680	8.83	27172	2.47
MFT602-S1	252.82	224194	191.59	17652	3.5	31855	2.04

*Including gaps coded as N

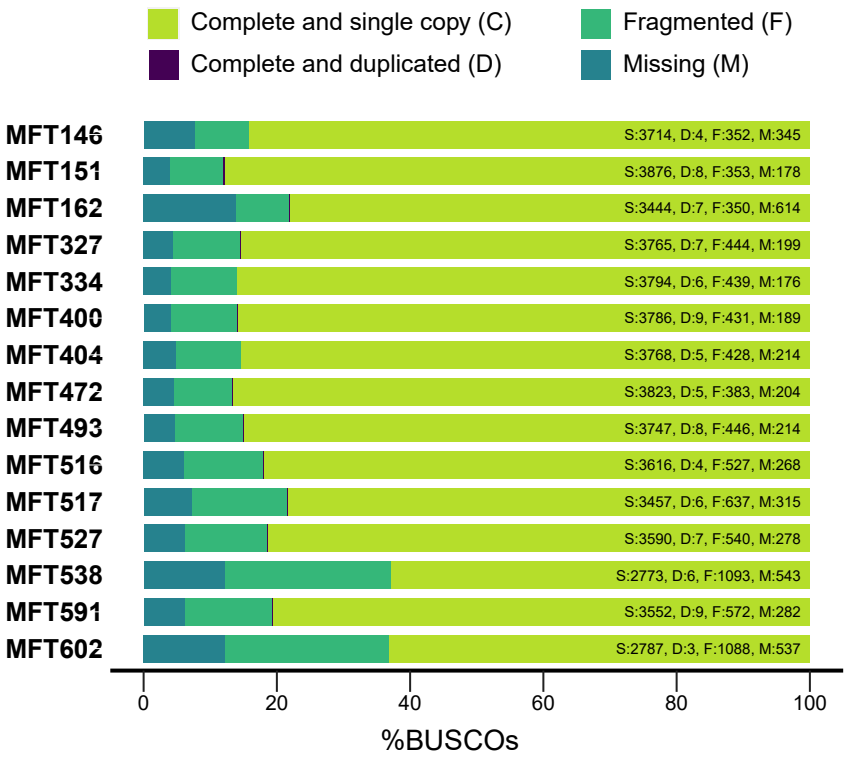


Figure 3.1: Percentage of complete, duplicated, fragmented and missing single-copy **BUSCO** genes for all *Azteca* assemblies, from a gene set of 4415 **BUSCO** genes in Hymenopterans.

3.3.2 Gene trees and species tree

From the **BUSCO** dataset of 4415 orthologues, 1412 single-copy genes are shared across all fifteen *Azteca* samples and 1395 across *Azteca* and *Linepithema*. The **ASTRAL** Maximum likelihood species tree of *Azteca* resulted in the same topology as the *Azteca* species tree with the outgroup *Linepithema*, with a local posterior probability close to one for all branches, but the number of proteins supporting each branch drops when moving from the internal (deeper) branches towards the tips (Figure 3.2). The **ASTRAL** local posterior probability is a function of the branch frequency given a number of genes, and a branch recovered, for instance, in 60% of all genes will still have

strong support if the number of total genes used to build the tree is large (Sayyari and Mirarab, 2016).

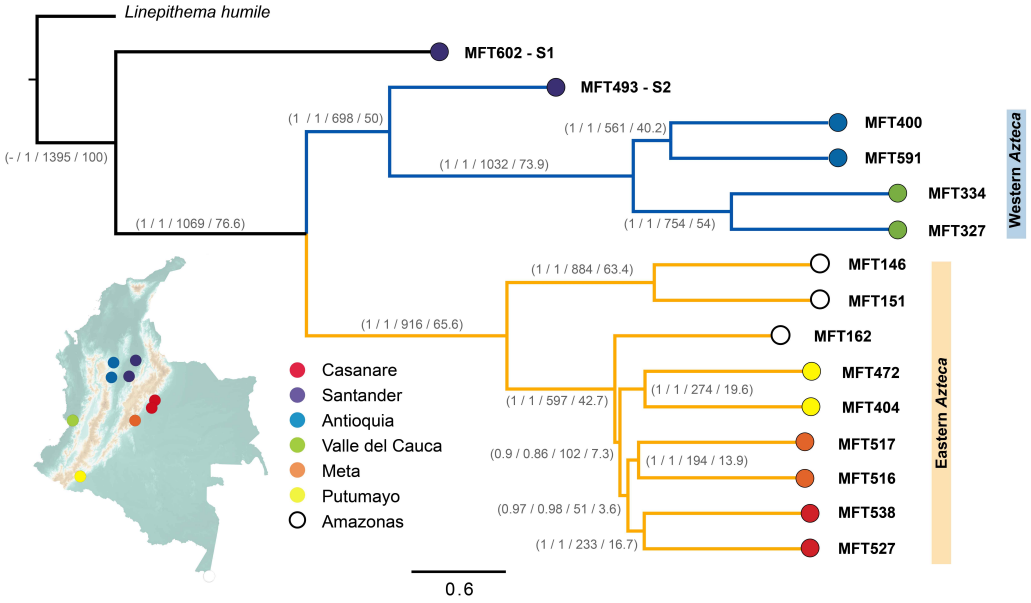


Figure 3.2: ASTRAL species tree from 1395 single-copy **BUSCO** genes shared across *Azteca* and the outgroup *Linepithema*. The values within brackets are from left to right: local posterior support for the unrooted tree, local posterior support for the rooted tree, number of **BUSCO** genes supporting that node, percentage of **BUSCO** genes (out of the 1395) supporting the node. Blue lines correspond to Western *Azteca* and orange lines correspond to Eastern *Azteca*. Samples are color coded as shown in the map.

Results from **Phyparts** showed that 76% of the genes support the placement of the S1 sample as sister to the remaining *Azteca*. This means that 76% of the gene trees have the same topology for that node, regardless of what the other nodes are for those gene trees. The remaining *Azteca* sequences are divided between strongly-supported Western and Eastern clades. Half the genes support S2 as part of the Western clade and an outgroup of the clade including Antioquia and Valle del Cauca samples, supported by 73% of the genes. However, 44.1% of genes support alternative topologies in which S2 belongs to the Eastern clade and 1.6% of the genes support S1 and S2 as sister lineages. The Eastern *Azteca* clade is supported by 65% of the genes with a clade of two Amazonian samples and a less supported clade including the rest of the Eastern samples. Branches

within this last group are significantly shorter compared with the rest of the tree and fewer genes support them, indicating a high number of conflicting topologies leading to those tips. Despite the high local posterior support in all branches, the number of supporting genes suggests conflicting gene topologies and potentially strong conflict between gene trees and the species tree (Figure C.2 in Appendix C). Very similar results were obtained for the Bayesian gene trees, with discordant topologies mainly involving the position of S2 (Figure 3.3). Moreover, both nuclear or mitochondrial-like single-copy genes show topological incongruences even with loci of their same genome.

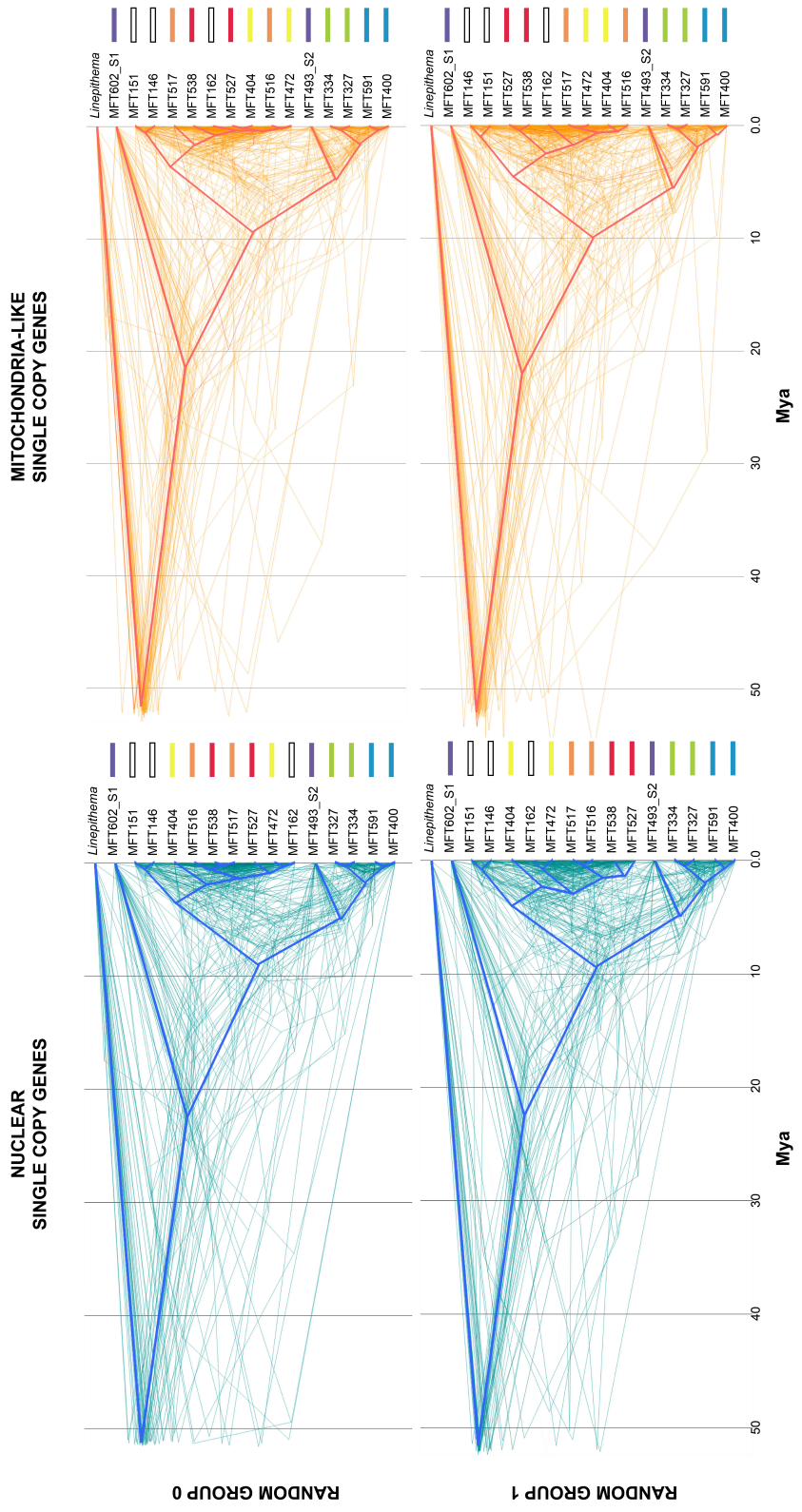


Figure 3.3: Bayesian gene trees and species trees of four sets of 50 randomly selected **BUSCO** genes. The species tree is represented by the thicker line and all samples are color coded as in Figure 3.2. The sets in blue represent nuclear genes and sets in orange represent mitochondria-like genes.

3.3.3 Gene tree age estimates

Calibrations based on Ward et al. (2010) resulted in dates and rate estimates that are consistent across most genes and independent of whether the genes are nuclear or mitochondria-like, with topologies that vary among genes. The split of the Andean lineage into Eastern and Western lineages is also consistently supported, although some genes have topologies in which the geographic structure does not hold (Figure 3.3). In the cases when alternative topologies support an east-west split and an Eastern membership of S2, the time for the most common ancestor between S2 and its closest relatives more often occurs between 0-5 Mya, after the Andean split (Figure 3.4). In general, discordance is much higher within Eastern and Western *Azteca*, but less so between them or between S1 and the rest of the clades.

When using fossil-derived calibrations, the times to the most recent ancestor (tMRCAs) and mean evolutionary rates were successfully estimated for all genes in the four datasets. Estimated mean rate is close to the values reported by Papadopoulou et al. (2010) and Ho and Lo (2013) for insects: 1.77% ±0.19% per My for mitochondrial loci and 1.84% ±1.52% for nuclear introns. However, the marginal probability distribution for the likelihood of species trees was bimodal and estimates of ancestral population sizes did not converge. The bimodal distribution is likely to be a consequence of the two most common alternative topologies involving the placement of S2. Nevertheless, the marginal probability distribution of gene trees, rates and calibrations always converged. Most estimated rates for nuclear and mitochondria-like genes are around 1.2% with a standard deviation of 3.28% per My, including outliers with unusually fast rate estimates (Table 3.3). Standard deviations reported here are wider than those in Papadopoulou et al. (2010) because of the larger number of loci sampled. It might also

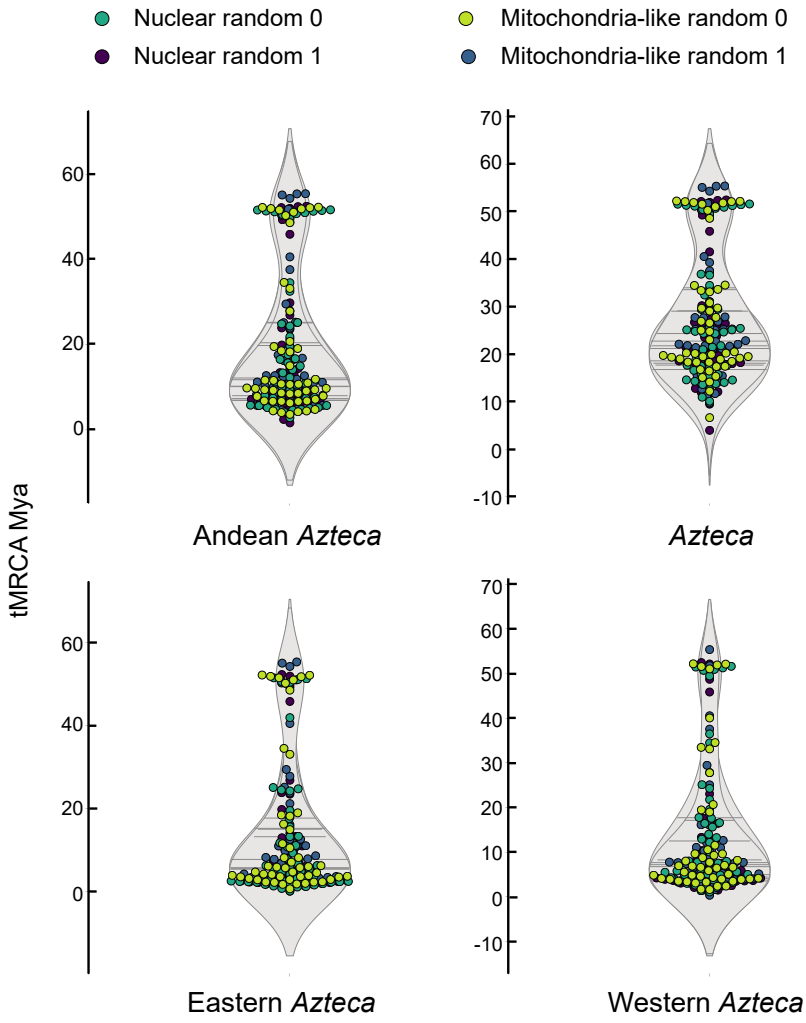


Figure 3.4: Violin plot of the tMRCAs for *Azteca*, Andean *Azteca* (or the split between the Western and Eastern lineages), Western and Eastern *Azteca* estimated using four sets of 50 randomly chosen genes.

result from different positions evolving at different rates; however, modelling partitions is a very limited option for the analysis of genomic data under Bayesian frameworks. When the rates reported by Papadopoulou et al. (2010) and Ho and Lo (2013) are used as priors for the rates, estimated rates are much faster and the variation across genes much wider. Such estimates vary between 2.57^{e-3}–4.6^{e-2} (mean= 7.57^{e-3}) for mtDNA and 6.33^{e-3} to 4.47^{e-2} (mean= 2.19^{e-2}) for nDNA, resulting in tMRCAs to *Azteca* between 1.2-11.12 Mya (mean= 4.9) for mtDNA and 0.55-19.9 Mya (mean= 2.41) for

nDNA. These results disagree with most studies on ant evolution and will not be shown here.

Node calibrations resulted in very similar age estimates for the four summary species trees with no difference between nuclear and mitochondria-like genes. The median tMRCA to crown *Azteca* included in this chapter is 22.3 Mya, when S1 diverges from the rest of *Azteca*, followed by a split of the Andean *Azteca* into the Eastern and Western lineages 10.95 Mya. At 5.9 Mya and 6.46 Mya respectively, the Eastern and Western+S2 lineages started diversifying (Table 3.3).

Most Bayesian gene trees support with posterior probabilities higher than 0.80 the Western and Eastern *Azteca* clades, independent of the genes being nuclear or mitochondria-like (Figure 3.5). In all datasets, more than 25 out of the 50 genes used support these two clades (Figure 3.6). The West+S2 is strongly supported by 25 or more gene trees; however, the distribution of posterior probabilities supporting the Western+S2 clade has a higher variation compared to those supporting either Western or Eastern *Azteca* alone. The East+S2 clade has a lower support, with posterior probabilities lower than 0.50 and 5 or fewer gene trees supporting the branch. Finally, the Santander (S1+S2) clade has not significant support.

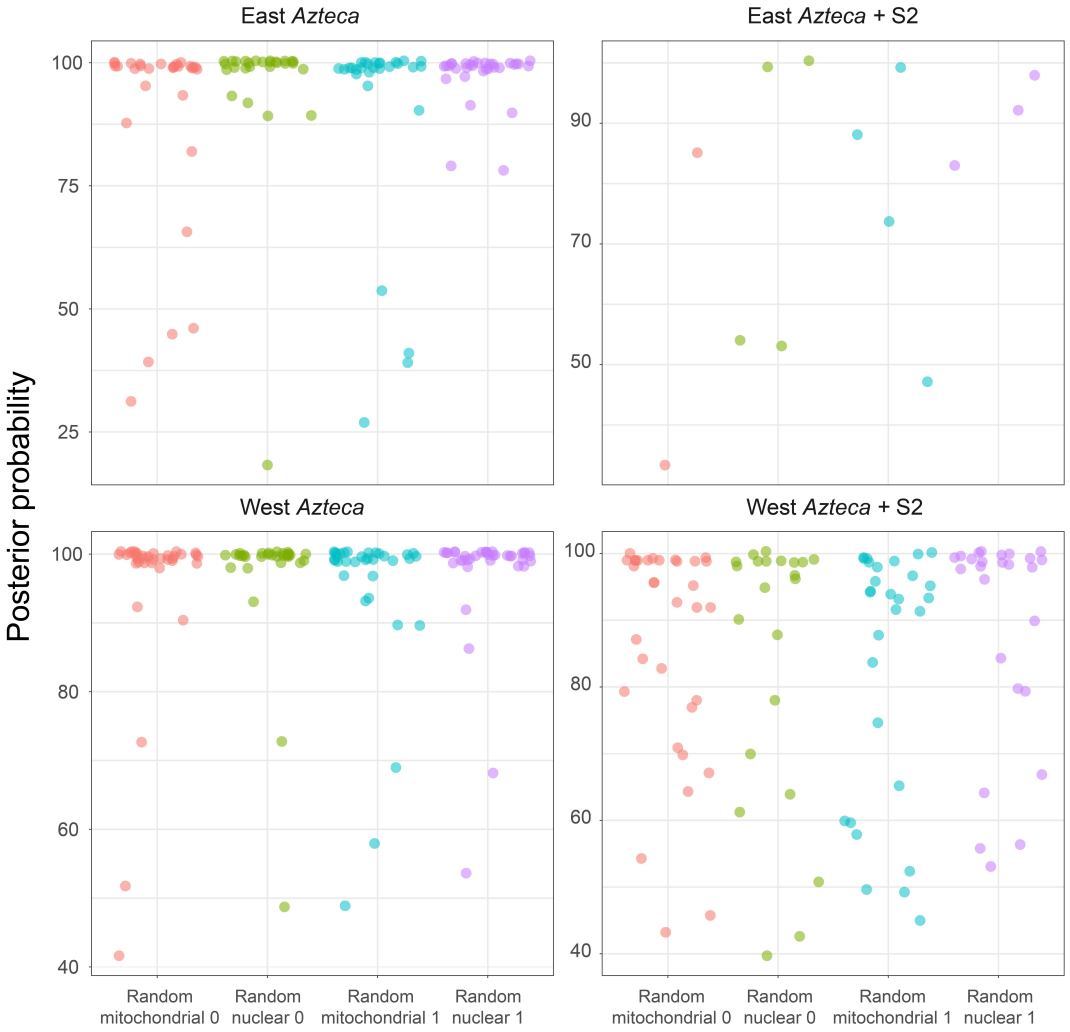


Figure 3.5: Branch posterior probabilities for *Azteca* clades including or excluding S2, derived from every gene tree estimated for the four nuclear and mitochondrial datasets. The posterior probability axis is expressed in percentage.

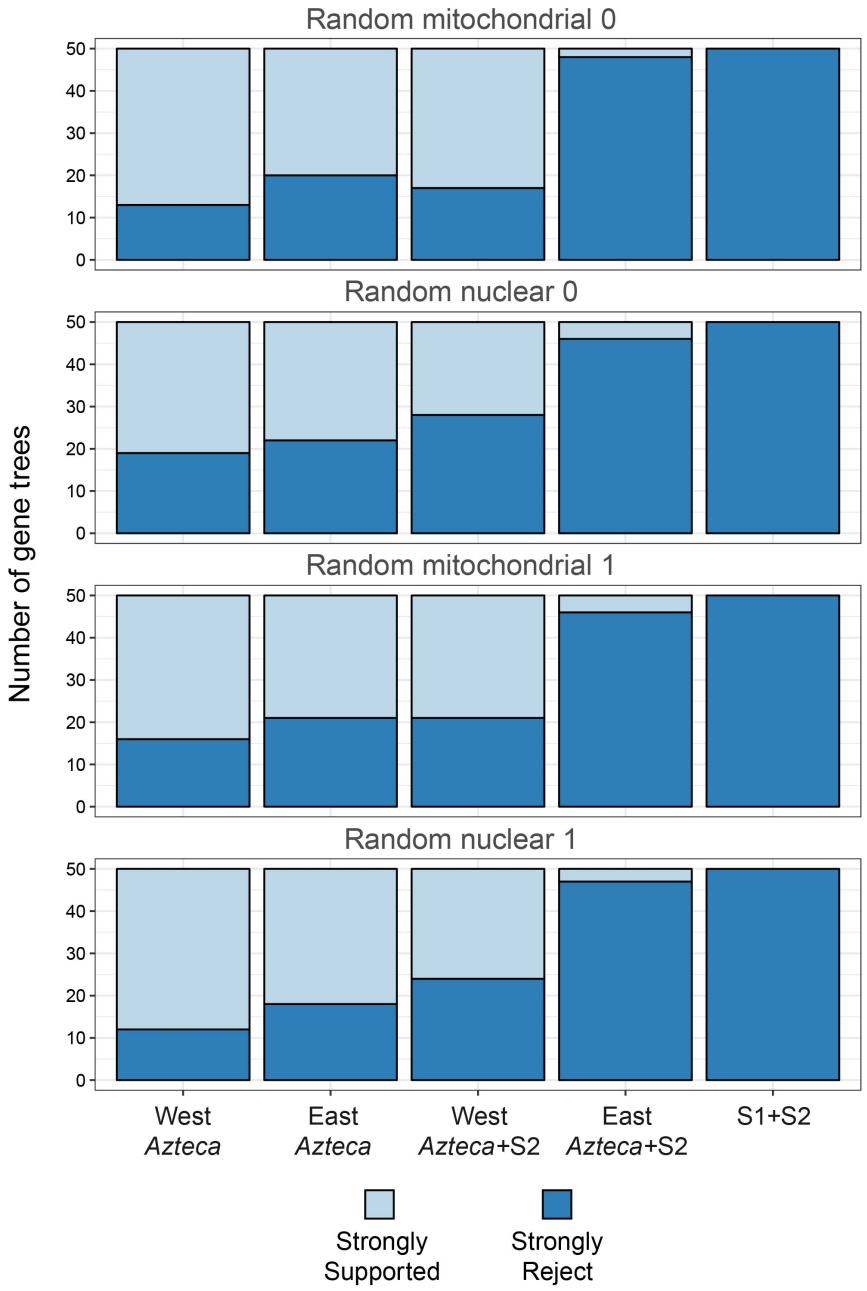


Figure 3.6: Number of gene trees from the nuclear and mitochondrial datasets whose topology and posterior probabilities strongly support or reject the different configurations of *Azteca* clades.

Table 3.3: Estimated rates and times to most recent common ancestor for the Andean *Azteca* lineages

		Time to most recent ancestor (Mya)		
	Estimated rate	**Andean *Azteca***	**Eastern *Azteca***	**Western *Azteca***
mean	1.80^{e-03}	18.73	13.18	12.46
median	1.02^{e-03}	10.95	5.9	6.46
SD	3.28^{e-03}	16.63	16	14.24
95% CI mean	8.95^{e-04} - 2.71^{e-03}	14.12 - 23.34	8.74 - 17.61	8.51 - 16.41
95% CI median	1.06^{e-04} - 1.92^{e-03}	6.34 - 15.56	1.47 - 10.34	2.51 - 10.41

3.3.4 *Wolbachia* species tree estimation

From a total of 57,043 contaminant contigs removed, half were identified as the alpha-Proteobacterium *Wolbachia*. Reads from those contigs were filtered from the corresponding host *Azteca* sample and a total of five *Wolbachia* genomes were assembled successfully, all of them from Western *Azteca* (Table C.1 in Appendix C). A total of 41 single-copy **BUSCO** genes (from a gene set of 221 **BUSCO** genes for Proteobacteria) are shared across the five *Wolbachia* and the reference *Wolbachia* genomes, and the **ASTRAL** tree from those genes group together the *Wolbachia* from *Azteca* in a clade within the **A** *Wolbachia* and sister to a clade of *Wolbachia* inhabiting *Drosophila* and *Nomada* bees (Figure 3.7).

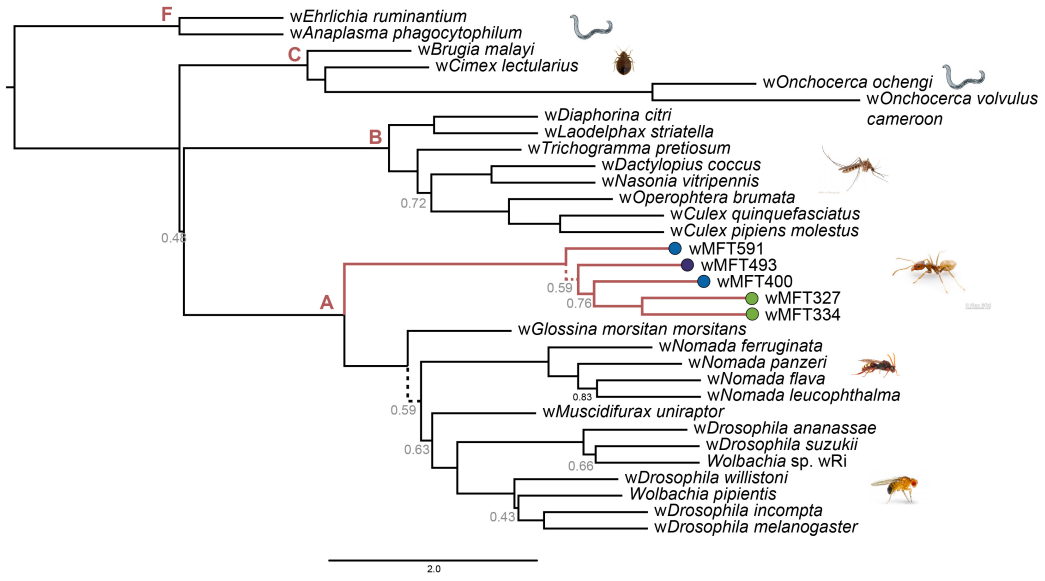


Figure 3.7: ASTRAL tree using 41 single-copy **BUSCO** genes shared across *Azteca-Wolbachia* and other *Wolbachia* reference genomes from NCBI. Red lines represent the *Wolbachia* strains assembled from the *Azteca* reads and all accessions are named with a "w" followed by the scientific name of the host, except for *Wolbachia pipientis*. Colors correspond to the geographic location of the samples: Antioquia in blue, Santander in purple and Valle del Cauca in green. Local posterior probabilities above 0.90 are not shown.

3.4 Discussion

3.4.1 *de novo* genome assembly

The resulting ant *de novo* genome assemblies are more complete than the *T. guianensis* assemblies, as discussed in Chapter 4. On average, *Azteca* assemblies have fewer duplications than *Tococa* assemblies and the percentages of fragmented and missing **BUSCO** genes are always less than 40% of the genome. More than 3400 single-copy **BUSCO** genes are complete in the best *Azteca* assemblies, compared to 1426 found in the *Aphaenogaster* (Myrmicinae) transcriptome (Stanton-Geddes et al., 2016). 125 genes showed best hits to mitochondrial proteins in Hymenoptera, compared to the 11 mitochondrial genes and 228 mitochondrial-like genes that were found in *Pseudomyrmex gracilis* genomes and 173 mitochondrial-like genes in *Atta cephalotes* (Rubin and Moreau, 2016). The fifteen *Azteca* assemblies were aligned, cleaned and variants were called following the GATK best practice pipeline, and they will be used in the future for more appropriate population analyses due to time limitations. For a low coverage *de novo* assembly of *Azteca* genomes, the results presented here are relevant. Future steps to improve the quality, coverage and completeness of the assemblies include a deeper coverage and the sequencing of fragments longer than 1000 base pairs.

3.4.2 Species tree estimation

Estimating gene and species trees of *Tococa*-associated *Azteca* populations allowed for the reconstruction of their phylogeographic history and unveiled processes that would not have been easy to detect using only two or three loci (see Chapter 4). Summary species trees from ML and Bayesian methods resulted in the same backbone with few

differences only in the relationships of samples within major clades. The main difference between the methods, however, is the computational scalability. Estimating ML trees gene by gene to later summarize them based on quartet topologies allows the use of hundreds of genes simultaneously while keeping computational requirements at a minimum. Bayesian approaches are advantageous because they account for uncertainty in the data; however, the analyses require weeks and are restricted in the number of genes that can be used. Moreover, both methods are robust to ILS, especially when the number of loci is high, but ML approaches are not well suited to calibration analyses and thus Bayesian methods are preferred.

3.4.3 Gene discordance and phylogeography

Despite the high support to the species tree topology, a large proportion of gene topologies are incongruent with the species trees, regardless of the gene being nuclear or mitochondrial. Under a scenario of mito-nuclear discordance caused by (for instance) mitochondrial capture, the expectations are that the mitochondrial and the nuclear summary species tree differ in topology and that all mitochondria-like gene trees will show similar topologies as the mitochondria does not recombine. Thus, the results show no preliminary evidence of mitochondrial capture causing the conflicts in the placement of S2. However, the incongruence observed among mitochondria-like genes is comparable to that among nuclear genes, even though mitochondrial genes are expected to provide better resolution as their effective population size is half that of nuclear genes.

Within Western and Eastern *Azteca* clades, the high discordance is consistent with within lineage gene flow between populations of *Azteca* located in the same side with respect to the Eastern Cordillera. This is particularly evident on all the different

topologies within Eastern *Azteca* no matter how geographically distant the populations are. The Meta and Valle del Cauca locations are closer to each other than Meta is to Amazonas or Putumayo, but genetically, Meta, Amazonas and Putumayo form a distinct clade from Valle del Cauca (Figure 3.2). Another explanation for the high gene discordance could be populations sizes. The bigger the effective population sizes are the more time it is required for the fixation of alleles and for the sorting of ancestral polymorphisms.

Topological incongruence between gene trees can result from a range of processes. Long generation times and large effective population sizes have the potential to increase incomplete lineage sorting. Admixture due to secondary contact after lineage divergence produces the same incongruence patterns and distinguishing this mechanism from simple ILS requires to compare topologies between pairs of populations with different gradients of overlap and at different distances from each other (Petit and Excoffier, 2009).

Two incipient species diverging in allopatry are estimated to require around 9-12 generations to achieve complete genetic differentiation in at least 95% of their loci, this is, to solve incomplete lineage sorting (Zhou et al., 2016). If these organisms are characterized by long generation times, then the process is longer in absolute time. Ant queens can have long life spans (up to 20 years for some species) that couple with continuous production of workers and fertile alates increases the overall generation times for a single ant colony (Keller, 1998). In polygynous species, more than one queen produces workers throughout the life of the colony, increasing the life span of the colony (Keller, 1998). In monogynous species, the single queen can be replaced by a new one after the first dies, also increasing the life span (DeHeer and Tscbinkel, 1998). The generation

time of an *Azteca* colony, particularly in the case of the species inhabiting plants, is still unknown and the only study available deals with the life span of *Azteca sp.* males (Shik and Kaspari, 2009). Thus, it is not possible to test the length of *Azteca*'s generation time as a cause of incomplete lineage sorting.

Effective population sizes for *Azteca* lineages were not estimated here, but future analyses using the genomic data derived from this study will be used to estimate the parameter. However, phylogenetic analyses in Chapter 2 suggest that both Eastern and Western *Azteca* lineages have large population sizes as populations within each clade lack nuclear geographic structure despite being distributed over large geographic areas. Consequently, it is possible that ILS is accentuated by continuous gene flow, large population sizes and simultaneous retention of ancestral polymorphisms, as will be discussed later.

In addition to incongruence within clades, incongruence between clades also occurs, although to a lesser degree. In a scenario of strict vicariance or allopatric speciation, all genes are expected to reflect the split at roughly the same time if there is no gene flow. But first, the effects of vicariance depend on the time scale, generation times and population sizes of the lineages and second, in some cases gene flow is likely to continue for a few generations allowing two isolating populations to share some alleles (Mallet, 2007). Moreover, speciation has different effects in different regions of the genome and those differences contribute to incongruent gene topologies (Beltrán et al., 2000; Leaché et al., 2016). The split between Eastern and Western *Azteca* occurred relatively recently and it is likely that not enough time has passed for the ancestral polymorphisms to be sorted. Evidence found in Chapter 2 of little gene flow between some populations across the Andes where the Cordillera is low in height suggests that

divergence occurred under continued gene flow at least for a while. This increases the time required for sorting the ancestral polymorphisms shared between Eastern and Western *Azteca*. Besides, large ancestral population sizes relative to the time between diverging events increase the probabilities of incomplete lineage sorting (Maddison and Knowles, 2006). Similar patterns of incomplete lineage sorting are observed in recently diverged species of Neotropical *Inga* plant (with all species arising between 2-10 Mya, Richardson et al. 2001) or in old rapid radiations like that in neo-avian birds during the Cretaceous-Paleogene (Jarvis et al., 2014).

Processes involving genome composition can also produce the incongruence patterns observed among *Azteca* genome assemblies. First, the introduction of mitochondrial genes into the nuclear genome or NUMTs (nuclear mitochondrial-like sequences) that are very similar to the functional mitochondrial copies but are under different evolutionary pressures that can mislead phylogenetic results (Sorenson and Quinn, 1998; Martins et al., 2007). It is possible that the mitochondria-like sequences from *Azteca* genome assemblies are unusually conserved NUMTs that the methods failed to recognize as nuclear copies of mitochondrial genes. If the mitochondrial copies are fragmented in the assembly and the NUMTs are complete and highly conserved (*e.g.* lacking stop codons in unusual positions) reciprocal blast will return the NUMTs as mitochondrial instead of nuclear. Under that scenario, loci identification using **BUSCO** can fail to report duplication. Thus, the gene trees from these loci can have significantly different topologies to those of mitochondrial copies generating patterns of incongruence that are not the result of ILS or other mechanisms. Second, as Yang (1998) says "neither too similar or too divergent molecular sequences contain much phylogenetic information". Saturation of substitutions occurs in fast evolving sequences as reverse mutations and homoplasies are more frequent than in slow evolving sequences. This translates in lost

information and failure to reconstruct the true topology (Xia et al., 2003), contributing to the incongruence between gene trees. To limit the noise from the results, it is necessary to use more stringent filters to eliminate undetected NUMTs and uninformative loci from the analyses.

3.4.4 Tree calibrations

Confidence intervals around the mean tMRCA to Andean *Azteca* using genomic data are narrower than those estimated in Chapter 2. From the ITS2 and COI estimations, the 95% confidence interval around the mean is 3.39-38.15 Mya across models and markers, but using genomic data the lower and upper bounds of the intervals are 14.12-23.34 Mya across nuclear and mitochondrial-like random sets of genes. Thus, including more genes reduces the variance around the mean estimates, at least in the case of this dataset. However, confidence intervals around the median suggest lower age estimates (Table 3.3), but as mentioned before, this is likely because of extreme values over the mean.

Gene trees and species tree age estimates provide evidence favoring the role of the Andean uplift in promoting population isolation and divergence and that such divergence took place under the presence of gene flow between some of the populations. The S1 population, located in Santander between the Eastern and the Central-Western Andean Cordilleras (Figure 2.1 in Chapter 2), is strongly supported as sister to the rest of *Azteca*. That, and the bifurcated divergence between Eastern and Western *Azteca* (*e.g.* no western specimens are nested within the Eastern lineage or vice-versa), suggests that both lineages derived from the split of a single ancestral population. That ancestral population could have been widely distributed around what is now the northern

Andes area, but extremes of that population became isolated as the mountains rose and eventually split (Figure 3.8). The Eastern Cordillera, although the highest and fastest rising of the three cordilleras, only closed completely 5.2 Mya by connecting its southern range with the Venezuelan range in the North of Colombia (near Santander) and closing the canyons generated by mountain building around 12 Mya (Hoorn et al., 2010). Currently, the Cordillera is a continuum with areas lower than 2,000 m.a.s.l. were the Colombian joins the Venezuelan Cordillera (Figure 5.1 in Chapter 5), near the sites were S1 and S2 were collected.

Regarding the conflicting position of S2, a likely scenario is that an ancestral population with the centroid (or origin) in Santander split into west and east by the rapid uplift of the Andes, originating the Western and Eastern *Azteca* clades. Then, the S2 lineage could have arisen as a lineage maintaining gene flow with S1 and Western *Azteca* via the northern lowlands and with Eastern *Azteca* via the gap in the Eastern Cordillera until it finally closed (arrows near 1 and 2 in Figure 5.1, Chapter 5). The proportion of gene tree topologies supporting the Western+S2 over the Eastern+S2 clade suggests that gene flow with Western *Azteca* continued for longer than with Eastern *Azteca* and that genes supporting the Eastern+S2 clade represent ancestral polymorphisms.

Gene flow between S2 and Eastern *Azteca* could have also stopped with the appearance of dry forest in the same area. The dry forest is a habitat were *Azteca* and *T. guianensis* have not been recorded (or observed, per. obs.), suggesting that this habitat could have contributed to the isolation of both populations. The fact that in topologies showing S2 as the sister (or outgroup) to the rest of Eastern *Azteca* and not nested within it supports this scenario. Similar cases of a porous Andean barrier are discussed in Chapter 2 and Chapter 4, between M3 and M4 *Azteca* populations and between *Tococa*

populations from Meta and Antioquia. Finally, the tMRCA to all Andean *Azteca* in the alternative topologies that lack geographic structure tends to be around or older than 10 Mya, suggesting that ILS and isolation with gene flow are the cause of the gene discordance, although secondary contact or more recent admixture cannot be completely ruled out.

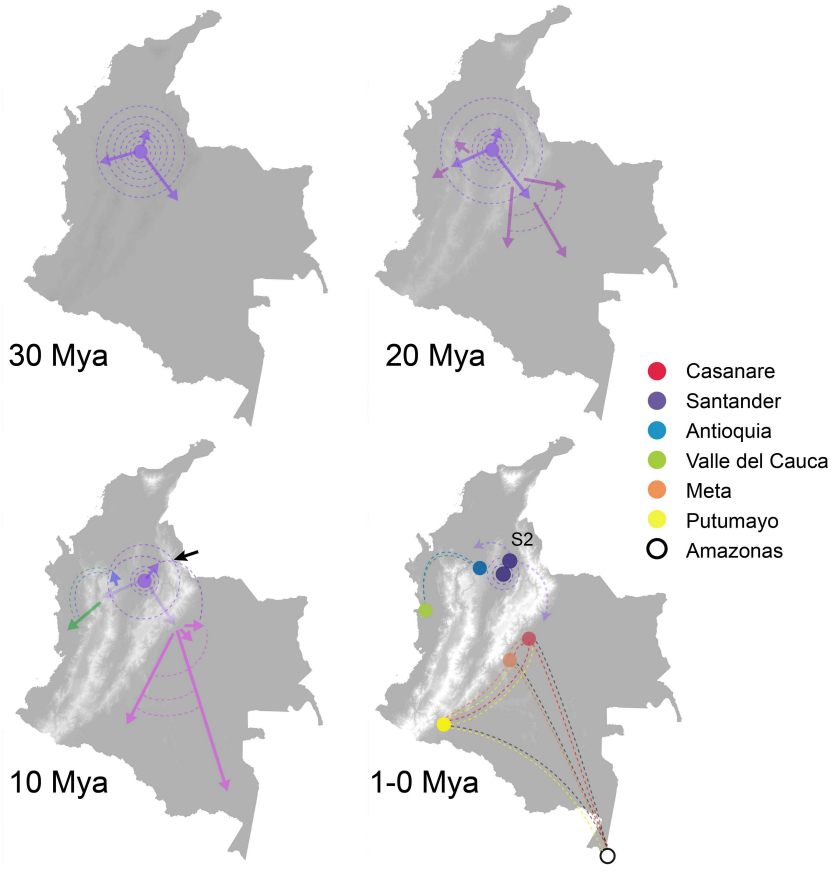


Figure 3.8: Hypothetical scenario for the diversification of *Tococa*-associated *Azteca*. An ancestral population was already present 30 Mya. This population either migrated or expanded across the territory around 20 and 10 Mya just before the Andean uplift isolated the eastern and western extremes of the population. Gene flow and population expansion are likely to have continued between Western and Santander and between Eastern and Santander (S2) populations until the final closure of the Eastern Cordillera 5.7 (marked with a black arrow). The final uplift around 2.7 Mya and the appearance of dry forest could have finally isolated the Western and Eastern *Azteca* lineages. Collecting locations of the samples are represented by filled circles and the colors represent the geographic location and the transition to new locations. Dashed circles represent areas where gene flow could have continued, and arrows represent directions of population expansion or bidirectional migration.

The placement of S1 as the outgroup to all *T. guianensis*-*Azteca* and better support for

S2 as the outgroup of Western *Azteca*, suggests that the scenario described above is more likely than a model of lineage migration from east to west followed by a population split. The latest scenario would have produced a phylogeny where the western populations were nested within the Santander population, both in turn nested within the eastern population. A similar result but in the opposite directions would be expected if the ancestral population originated in the west. However, the data does not support those two hypotheses. Future estimation of ancestral population sizes will provide information on population expansions or bottlenecks, but currently, the individuals sequenced for the Santander populations are not sufficient in number for a reliable inference.

3.4.5 *Wolbachia* symbionts

All *Wolbachia* symbionts successfully assembled belong to Western *Azteca*, and neither the absence of *Wolbachia* in Eastern *Azteca* or the lack of coverage for those samples can be ruled out. Assuming all *Wolbachia* genomes could have been assembled, a comparative analysis between the ant and symbiont phylogenies would have revealed either coevolution between *Wolbachia* and *Azteca*, *e.g.* mediated by reproduction control on the ants by *Wolbachia* or vicariance in *Wolbachia* due to vertical transmission. An equally interesting hypothetical scenario is that *Wolbachia* has only infected Western *Azteca*, which together with its absence from Eastern *Azteca*, promoted the divergence between both lineages.

This chapter represents the production of 15 *Azteca* and 5 *Wolbachia* genome assemblies from which I estimated gene and species trees. The incongruence between mitochondrial and nuclear genomes was not satisfactorily explained by mitochondrial capture, but rather by ILS. Similarly, the conflicting position of S1 and S2 with respect to the

rest of *Azteca* is not only explained by mito-nuclear incongruence. Further analyses incorporating appropriate parameter estimations and model testing are needed to elucidate in more detail the mechanisms behind ILS. Such model testing at the population level allows for the detection of migration, possible recent admixture or hybridization. Here I also demonstrate how incorporating more loci into tMRCA estimations narrowed the certainty range around the time to divergence between Eastern and Western *Azteca*. However, increasing the number of loci for species tree estimation and calibration reveals intricate patterns of evolution that are not always resolved and often conflict. Nevertheless, it provides a more complete picture of the species history and consensus estimates that are less biased by individual gene histories. Genomic data provides good resolution and enough information to resolve relationships under ILS; however, it failed to resolve short branches that have not accumulated enough changes (typical of fast radiations or recent lineage divergences). Finally, even though a comparison between the horizontally transmitted *T. guianensis-Azteca* mutualism and the vertically transmitted *Wolbachia-Azteca* one was not possible, here I demonstrate that is possible to recover the *Wolbachia* symbiont genomes from ant hosts.

CHAPTER

4

PHYLOGENY AND BIOGEOGRAPHY OF THE *TOCOCA GUIANENSIS* GROUP AND ITS *DE NOVO* GENOME ASSEMBLY

4.1 Introduction

Tococa belongs to the Miconieae tribe in Melastomataceae and includes myrmecophyte and non-myrmecophyte species that exhibit large morphological variations that makes

morphological identification challenging (Michelangeli, 2005, 2010a). This is not unique to *Tococa*, it is also observed in other Miconieae genera like *Clidemia, Miconia* and *Conostegia*. Miconieae is the largest of the 20 tribes within Melastomataceae and according to Ocampo et al. (2014) includes over 1,850 species and 17 genera restricted to the Neotropics. Within the tribe, morphological characters not always correspond to the classification of species (Ocampo et al., 2014), and only one out of the 17 genera are monophyletic (Michelangeli et al., 2004; Goldenberg et al., 2008; Martin et al., 2008; Michelangeli et al., 2008; Ocampo et al., 2014). Fast diversification rates are also a characteristic of the tribe. In a study encompassing the order Myrtales (including Melastomataceae), Berger et al. (2016) estimated that the net diversification rate in Melastomataceae is seven times higher than the average in Myrtales. Their results support an increment of the rate occurring in the branch leading to *Clidemia, Miconia*, and *Tococa*. High diversification rate, morphological diversity and high number of species are features observed in rapid radiations.

Resolving phylogenetic relationships between rapidly radiating taxa is challenging, especially when the radiation occurred recently. This occurs because radiations result in an increased variety of diverging organisms (Linder, 2008) in a period of time that is not enough to accumulate genetic differences (Richardson et al., 2001). Neotropical plants like *Lupinus*, *Gentienella*, *Inga*, and *Guatteria* are examples of rapid and recent radiations triggered by changes in geography, habitat and climatic changes (von Hagen and Kadereit, 2001; Richardson et al., 2001; Hughes and Eastwood, 2006). What these genera and *Tococa* have in common is having a large number of species, recent diversification events, and conflicting species relationships (low sequence divergence between taxa, lack of resolution and support for monophyly) (von Hagen and Kadereit, 2001; Bell and Donoghue, 2005; Hughes and Eastwood, 2006). As rapid diversifications

are characterized by few nucleotide substitutions separating species in a phylogeny, the low information content and short branch lengths makes it difficult to resolve the relationships between species.

Traditional molecular markers and barcodes are often used to solve species relationships in plants. However, the performance of DNA barcodes to discern species relationships tends to be lower in plants than in animals (90% versus 70% of identified species are monophyletic as determined using DNA barcodes). This could be due to lack of genetic variation among closely related plant species, hybridization and/or gene tree paraphyly (Fazekas et al., 2009). Moreover, resolution varies across taxonomic levels. For example, analyses using the chloroplast matK gene can identify diverging orchid species as monophyletic in 90% of the cases (Kress and Erickson, 2007). It is also useful to confirm the monophyly of distantly diverged species of angiosperms in 92% of the cases (Lahaye et al., 2008; CBOL et al., 2009). But matK performs poorly and has less resolution among more closely related taxa within families like Myristicaceae (Newmaster et al., 2008) and other Orchid subfamilies (Neubig et al., 2009).

Contrary to the use of a limited number of loci, the availability of genomic sequences provides a better picture of the evolutionary history of rapidly diverging plant families, especially when information from multiple gene histories is integrated (Irwin, 2002; Shaffer et al., 2007). Genomic data has become popular as a resource to address conflicting phylogenetic relationships among highly diverse taxa and to identify informative markers and increase the resolution in phylogenies. For instance, estimating the frequency of hundreds of single nucleotide polymorphisms (SNPs) across populations, and the detection of loci under selection, are applications made possible when genome sequences are available (Kumar et al., 2012; Olsson et al., 2017).

The phylogeny of Miconieae is not well resolved as it includes rapidly and recently diverged groups and paraphyletic genera and species, including *T. guianensis* (Michelangeli et al., 2004). *Tococa* itself includes 45 officially recognized species and other 18 unresolved taxa (Michelangeli, 2005, and The Plant List `http://www.theplantlist.org/tpl1.1/search?q=tococa`, accessed on Feb. 2nd, 2018). Limited genetic and genomic resources for these and related species limit the ability to resolve evolutionary relationships between species and to identify taxa using DNA barcoding. To understand the dynamics of the *Tococa*-ant mutualism, it is important to identify the taxa involved and to address the phylogenetic relationships among hosts. This can be a complex task, and the first step to achieve it is to generate useful genomic data. Thus, the aims of this chapter are to identify sampled *Tococa* specimens using molecular markers and estimate the phylogenetic placement of the samples with respect to *T. guianensis* and its closely related species. Then, the aim is to describe the sequencing of *de novo* draft genome assemblies for *T. guianensis* specimens and to outline a strategy for identification of new genetic resources for phylogenetic analysis and molecular taxonomy.

4.1.1 Next Generation Sequencing

New technologies for DNA and RNA sequencing are opening the door to more exhaustive, informative studies that produce large quantities of molecular data at an affordable price. Next-generation sequencing (NGS) methods make possible the sequencing of whole genomes or transcriptomes from non-model organisms in less time than it takes to standardize a protocol for sequencing several loci using the Sanger approach. Different NGS technologies can produce short (25-700 bp) or long reads (10-20 Kb), single end or paired-end, that are subsequently processed and assembled with the

help of bioinformatics tools (Unamba et al., 2015). Most assemblers build a graph (De Bruijn graph) of reads aligned by their overlapping segments that are collapsed into contigs and then into scaffolds until as much of the genome length as possible is recovered. The more reads overlap over the same area, the deeper the coverage it has and the more confidence there is to deal with repeats and potential error. In an ideal scenario, genomes will have enough unique patterns such that each read can be merged to its original neighboring reads and its position will not be ambiguous. Previously assembled, curated and annotated genomes from closely related species can be integrated to help the algorithm to solve conflicting read positions. However, those resources are still limited. Obtaining the complete assembly of large and complex genomes is challenging and a low number of genomes (typically from model organisms) have been confidently assembled. Only 61 reference genomes of plant species were available in 2015, out of 245 eukaryote genomes (O'Leary et al., 2016). Even though obtaining raw genomic data is increasingly easy, methodological limitations associated with processing them can impose a bottleneck when working with certain organisms.

Limitations are caused by either the nature of the genome or by systematic errors introduced during sequencing. Problems associated with the nature of the genome are more pronounced in plants than in animals for several reasons. First, plants have chloroplast in addition to mitochondrial organellar genomes. Both have higher coverage than nuclear reads because of their higher copy number, resulting in lower coverage for nuclear loci when all templates are sequenced together, compared to animals where only the mitochondria interfere. Therefore, separating correctly reads for homologous genes present in both organellar and nuclear genomes can be challenging (and error-prone) for genome assembly, especially when the number of reads and the coverage is low (Claros et al., 2012). Second, whilst some plant genomes can be conveniently small (as far as

required effort to sequence them is concerned), others are larger by several orders of magnitude, with C-values ranging from the small *Genlisea margaretae* genome (Lentibulariaceae, with approximately 63 Mbp or 0.06 pg, Greilhuber et al. 2006) to the large *Paris japonica* genome (Melianthaceae with 148851.6 Mbp or 152.20 pg, Pellicer et al. 2010). Third, plants have a higher rate of transposable element (TE) accumulation, accounting for approximately two-thirds of variable genetic material (Lisch, 2013; El Baidouri et al., 2014; Elliott and Gregory, 2015). Their abundance and repetitive nature add more obstacles to the process of genome assembly (Schatz et al., 2010; Claros et al., 2012). More complications are added by the presence of paralogous genes that resulted from at least two ancient whole genome duplications (palaeopolyploidization) that took place soon after the origin and radiation of the angiosperms (Simillion et al., 2002; Bodt et al., 2005; Jiao et al., 2011). Paralogous genes are surrounded by repetitive transposons and pseudogenes resulting from the low selective pressure and physical closeness on the chromosome (Freeling et al., 2008). Resolving the position of the functional copy and paralogue(s) of a given gene when both copies are only slightly different is a complex process and reads from both can be merged in an artefact single "gene" causing the omission of the real gene from the final assembly (Claros et al., 2012). Genome duplications and TE prevalence are positively correlated with genome size and prevalence of repeats (Elliott and Gregory, 2015), contributing to the complexity of plant genomes and resulting in higher ploidy and heterozygosity than in other organisms (Meyers and Levin, 2006; Gore et al., 2009; Schatz et al., 2012). While these features might to confer evolutionary advantages like gene regulation, transcription factors, and multiple sets of metabolites producing genes (Bodt et al., 2005; Zahn et al., 2005; Elliott and Gregory, 2015), they contribute to the complexity of assembling plant genomes (Claros et al., 2012).

Systematic errors of sequencing technologies pose another limit to genome assemblers. Library preparation protocols including PCR steps are more likely to introduce artefacts that are easily mistaken as true site variants than those without PCR steps. Contamination during library preparation and sequencing introduces foreign DNA and modifies the coverage and GC% content of the reads increasing the ambiguities the assemblers must resolve. Contamination with fungi and bacteria can also add genes that are not originally present in the target organisms. The high profile misassembled tardigrade genome illustrates the incorrect assumptions made in final assemblies when contamination is not removed from the reads (Koutsovoulos et al., 2016). Another inherited problem with sequencing is related with read length. Read length influences the resolution and error rate since there is more information bridging repeats and variable regions in longer reads, but shorter reads have less size variance and the nucleotide calls are more reliable (Schatz et al., 2012). Additionally, different technologies have different error rates per nucleotide that can be up to 10% for Illumina and Roche 454 or even up to 15% for single-molecule sequencing such as Nanopore (Paszkiewicz and Studholme, 2010; Rasko et al., 2011; Unamba et al., 2015) Miscalled nucleotides, sites or indels introduce errors into the final assembly affecting downstream analyses and leading to incorrect biological inferences (Florea et al., 2011; Schatz et al., 2012). Nevertheless, systematic errors are more tractable than TE related errors, repetitions or genome size as error rates can be modelled and tracked. Additional software to filter reads before and after the assembly stage and protocols to estimate the quality of assemblies have been developed and can significantly improve resulting genomes. Despite the challenge of assembling plant genomes, new tools and an understanding of the evolutionary processes shaping their genomes make it possible to achieve assemblies of suitable quality for downstream studies.

4.1.2 *Tococa* plants

The origin of the Melastomataceae family is likely in Gondwana around 84-88 Mya (Morley and Dick, 2003). Based on plate tectonic separation and one Myrtaceae fossil, rate calibrations of the ndhF chloroplast marker suggest that the sister tribes Merianeae and Miconieae diverged around 65 Mya. Then Miconieae became fully established in South America by 55 Mya (Morley and Dick, 2003). Regarding Melastomataceae diversity, Goldenberg et al. (2008) have suggested a strong positive correlation between geography and the diversification of the tribe and genera within it: closely related species are distributed in geographically close areas.

Tococa is a Neotropical genus of shrubs and small trees with 45 recognized species (Michelangeli, 2005). First described by Aublet (1775) (who also described *Maieta*) from a *T. guianensis* type from French Guyana as a genus of myrmecophyte plants. The first genus monograph was written by de Candolle (1828), who included 5 more *Tococa* species and described other two non-myrmecophytes as *Miconia* and *Truncaria* based on the lack of domatia. These non-myrmecophyte species are now part of *Tococa*. With new species added to the genus by Martius (1832); Bentham (1840, 1844) and Bentham (1845), Bentham (1840) divided *Tococa* into three sections based on the presence and shape of domatia. This system recognized non-myrmecophyte plants as *Tococa* species for the first time. That classification was accepted until Naudin (1851) classified again Bentham's three sections into two sections based on the shape of the hypanthium, a non-myrmecophyte trait. According to Michelangeli (2005), Naudin's classification potentially placed closely related species into different sections, complicating more the taxonomic classification of the genus. After Naudin's work, Triana

(1871) organized the specimens collected by Spruce throughout South America, described new species from those and included even more non-myrmecophyte in *Tococa* based on morphological resemblance and dismissing the lack of domatia. He also reinstated the classification proposed by Bentham (1840). Both Naudin's and Bentham's classification were incorporated by Cogniaux (1891) who recognized a total of 38 *Tococa* species. Since then, the latest comprehensive work to describe *Tococa* is summarized in the monograph published by Michelangeli (2005). Of the species accepted and classified throughout *Tococa*'s systematics history, *T. guianensis* has represented the genus type and has not changed since the time the species was first described.

The genus is mainly Amazonian, though the most widespread species *T. guianensis* reaches north to southern Mexico, Belize and the Antilles (Michelangeli, 2005). Distributed from zero to 3000 meters above sea level (m.a.s.l.), plants from this genus become fertile early during their development (Michelangeli, 2010b). *Tococa* produces berries with abundant pulp that are most likely dispersed by birds and mammals. Around 30 out of the 45 *Tococa* species recognized have domatia for ants located either at the apex of the petiole or at the base of the leaf blade (Michelangeli, 2005, 2010a). The shape of domatia, leaves and the density and type of trichomes are traits commonly used to describe and classify specimens, sometimes unaware of the continuum variation of these characters among and within species (Michelangeli, 2005). Although most Miconieae myrmecophytes belong to *Tococa*, other genera like *Conostegia* and *Clidemia* also bear ant domatia (Michelangeli, 2010a). Because of extensive morphological variation, *Tococa* has a complicated taxonomy and the relationships to other genera in the tribe are unclear.

4.1.3 Previous phylogenetic analyses of *Tococa*

The circumscription of Miconieae is based on fruit and seed morphology and additional molecular data. Even though the relationships of the tribe with other Melastomataceae tribes is well established, relationships within it concerning mainly the *Tococa*, *Leandria*, *Maieta* and *Clidemia* genera are confusing. Moreover, the monophyly of individual genera has not been confirmed. Various attempts to resolve the relations within Miconieae have included some species of *Tococa* (Michelangeli et al., 2004; Goldenberg et al., 2008), although the phylogenetic reconstruction with the most complete species sampling used morphological but not molecular traits (Michelangeli, 2000).

The first attempt to resolve species relationships within *Tococa* was a cladistic analyses of approximately 60 morphological characters described from habitat, the development and shape of seeds, stem, domatia, leaf, inflorescence and infructescence characters (Michelangeli, 2000). Using 42 *Tococa* and 11 *Miconia* species, the resulting cladogram showed two non-monophyletic clades embedded within *Tococa*, defined mostly by the shape of the seeds. Subsequent efforts were focused on resolving species relationships within Miconieae including several (but not all) *Tococa* species. Michelangeli et al. (2004) reconstructed a phylogeny of the tribe based on nuclear ITS sequence data, including 15 *Tococa* species. This resulted in a well-supported core clade (*Tococa sensu stricto*, the clade where must *Tococa* species belong) with non-myrmecophyte *Tococa* at the base and a derived group of domatia-bearing *Tococa* species. The two non-myrmecophyte *T. broadwayi* and *T. perclara* fall in a separate clade with other species of *Tococa*, *Miconia* and *Clidemia* (*Tococa sensu lato*). *T. caquetana* falls in a more distantly related clade composed of *Clidemia* and *Necranium* (Michelangeli et al., 2004). In a later study, chloroplast *ndhF* sequence data and more species (including one *Tococa*

accession) were added to the Michelangeli et al. (2004) dataset only to confirm *Tococa* polyphyly (Michelangeli et al., 2008). In a phylogenetic evaluation of the Miconieae genus *Leandra* and its sister genera, petal and seed morphology were used as morphological characters along with ITS sequence data, placing the two non-myrmecophyte *Leandra* species with *Tococa*, *Clidemia*, *Anaectocalyx* and *Mecranium* (Martin et al., 2008). The relationships between this clade and the core of *Tococa* are unresolved in a polytomy that includes the majority of the remaining *Tococa*, *Leandra* and Clidemia species. Finally, a phylogenetic reconstruction of the paraphyletic *Tococa* (19 species) and other genera within Miconieae using ITS and ndhF confirmed the placement of the core *Tococa* as sister clade to Caribbean *Tococa* + *Conostegia* + *T. spadiciflora*, *T. broadwayi* and *T. perclara* within a *Mecranium* + *Anaectocalyx* + allies clade, and *T. caquetana* within the distantly related *Clidemia* clade (Goldenberg et al., 2008). Previous efforts failed to prove reciprocal monophyly of *Tococa* despite the sampling and the use of morphological characters; however, *Tococa sensu lato* and *Tococa sensu stricto* are repeatedly recovered, specially by ITS.

None of these studies tested the performance of ITS within species but the consistency with which recovers two distinct groups of *Tococa* species might be useful for specimen identification. Once the samples are identified as belonging to *Tococa sensu stricto* and as close relatives to reference *T. guianensis* sequences, specimens can be selected for further genomic analyses and marker development. Screening for more variable regions and developing new markers to reconstruct phylogenies would be greatly aided by a reference genome assembly. Likewise, further studies involving the plant (*e.g.* species delimitation within *Tococa*, population genetics and dynamics, genome-wide association studies for analysis of quantitative traits) would be boosted by the availability of the first Melastomataceae whole genome sequence and assembly.

The first aim of this chapter is to use molecular markers to identify specimens of *Tococa* cf. *guianensis* collected in Colombia, evaluate their position in relation to other specimens and Miconieae reference sequences, and select samples identified as *T. guianensis* for further genome sequencing. Molecular identification has advantages over morphological identification since it eliminates the need for fertile structures (*i.e.* diagnostic traits) and provides a useful background for comparisons when reference sequences are available. I also assess the extent to which the loci used in taxon definition show geographic structuring within and among lineages, particularly with respect to the Andes Cordillera. Geographic structure and genetic distances between *T. guianensis* populations provide insight on whether the Andes acted as a barrier to gene flow. This provides a basis for comparison between the phylogeographic patterns of *T. guianensis* and associated *Azteca* ants that will be discussed in Chapter 5.

The second aim of this chapter is to produce a *de novo* genome for *T. guianensis* useful as a reference for population analyses. Two transcriptomes of related taxa (*Tetrazygia bicolor* and *Medinilla magnifica*, are available upon request from the OneKP project (`https://sites.google.com/a/ualberta.ca/onekp/home`). However, the information that can be extracted from a transcriptome is limited to coding regions (exomes). Introns or non-coding regions, missing in transcriptomes, usually have higher evolutionary rates and provide more information on demographic processes such as phylogeographic divergence or changes in population size. Recently, individual phylogenies were reconstructed for sets of coding, non-coding and coding+non-coding regions in the plastid genomes of 16 Melastomataceae species, including *Miconia dodecandra* (Reginato and Michelangeli, 2016). However, their phylogenetic reconstructions using non-coding regions resulted in different topologies to those using coding and coding+non-coding regions. Furthermore, they report topological incongruence between their rbcL, ndhF

and rpl16 intron phylogenies and their coding+non-coding phylogeny. Regarding reference genomes available, *Eucalyptus grandis* (Myrtaceae) is the most closely related whole genome sequence (Myburg et al., 2014), but is genetically distant from *Tococa*. Assembling the first whole genome for the family is an important step to address the evolutionary relationships of one of the most diverse families of plants. From the genomic data, I show the potential for identification of phylogenetically informative markers that examining loci with different mutation rates has to resolve relationships between four accessions of *T. guianensis*.

4.2 Methods

4.2.1 Plant collections

Plants were sampled in the areas indicated in Figure 2.1 in Chapter 2. Locations were selected based on herbarium records of *T. guianensis*. Other places with similar environmental conditions were explored in an attempt to find unreported populations. A leaf from each plant sampled was collected in silica gel, choosing the healthiest and cleanest leaf available. Herbarium samples from fertile plants were pressed, dried and deposited in the Herbario Forestal UDBC -Universidad Distrital Francisco Jose de Caldas, Bogota-Colombia. Plant collections were made under the Macro permit granted by the National Authority for Environment Licenses, Colombia to the Universidad Distrital. Other representatives of Melastomataceae were collected at areas where no myrmecophyte *Tococa* plants were found.

4.2.2 DNA extraction and Sanger sequencing

Tococa cf. *guianensis* samples from different populations were selected to carry out Sanger sequencing of nuclear and plastid regions to confirm species identity, select accessions for genome sequencing, and assess how informative sequenced loci are at identifying potential genetic diversity. DNA extraction in Melastomataceae is challenging because they contain a high concentration of secondary products (Renner et al. 2001, Michelangeli pers. comm. and personal experience). Thus, DNA extraction from *Tococa* was performed using the Qiagen Plant DNeasy kit with the following modifications. Veins were removed from the silica dried leaves and a piece of approximately 1x3 cm was macerated on the TissueLyser II (Qiagen, Germany) for two minutes at 20 Hz. 400μL of AP1 buffer, 30μL Proteinase K (10mg/mL) and 30μL β-Mercaptoethanol were added to each sample, which was then left incubating overnight at 65°C. During the final elution step, highly degraded DNA fragments were removed from the column by adding 50μL of EB buffer followed by immediate centrifugation. To recover as much DNA as possible, DNA was eluted from the column by applying 40μL of EB buffer, then the flow-through was loaded back to the column, incubated and centrifuged. A final elution was performed with 25μL.

Two chloroplast (ndhF and ycf1b) and one nuclear (ITS) regions were amplified and sequenced (Table 4.1). The nuclear and chloroplast molecular markers were first tested using Sanger sequencing on a subset of specimens from all populations before fully committing with more time and reagents. The PCR mix for ITS and ndhF was as follows: 1μL of template DNA was added to a final volume of 20μL containing 0.16mM dNTPs mix, 1x PCR Buffer, 2.25mM MgCl2, 2μM of each primer, 10μg/μL BSA and 0.3 units of Taq (Bioline). PCR mix for ycf1b differed by having 0.25mM dNTPs

mix, 2.5mM MgCl2, 1μM of each primer, and no BSA. Cycling conditions for ndhF were 5 min at 94°C followed by 10 cycles of 10 sec at 94°C, 45 sec at 45°C, 50 sec at 72°C; followed by 25 cycles of 10 sec at 94°C, 45 sec at 48°C, 50 sec 72°C, with a final extension of 10 min at 72°C. Cycling conditions for ITS were 2 min at 94°C followed by 35 cycles of 10 sec at 94°C, 45 sec at 50°C, 50 sec at 72°C and with a final extension of 10 min at 72°C. Cycling conditions for ycf1b were 4 min at 94°C followed by 34 cycles of 30 sec at 94°C, 40 sec at 50°C, 1 min at 72°C and with a final extension of 10 min at 72°C. PCR products were visualized in a 2% agarose gel stained with SYBRGreen, then cleaned following the shrimp alkaline phosphatase and exonuclease I protocol and subsequently sequenced in both directions on an ABI 3730 capillary machine using BigDye version 3.1 terminator chemistry (Applied Biosystems). Sequences were aligned using the **MUSCLE** algorithm (Edgar, 2004) implemented in **Geneious v.4.8.5** (`http://www.geneious.com`, Kearse et al. 2012), then quality checked and edited by eye.

Table 4.1: Primers for the amplification of chloroplast and nuclear regions from *T. guianensis* accessions

Region	Genome	Primers	Sequence	Reference	Length
ndhF	Chloroplast	ndhF-1318F	GGATTAACCTGCATTTTATATGTTTCG	(1)	616 bp
		ndhF-1955R	CGATTATAT GACCAATCATATA		
		ndhF-972F	GTCTCAATTGGGTTATATGATG	(1)	
		ndhF-1603R	GCATAGTATTGTCCGATTCATAGAGG		
ycf1b	Chloroplast	ycf1bF-Myrtaceae	TCTCGACGAAAATCTGATTGTTGTGAAT	(2)	777 bp
		ycf1bR-Myrtaceae	ATATATGTCGAAACAATGGAAA		
ITS1	Nuclear	ITS8P	CACGCTTCTCCAGACTACA	(3)	798 bp

Table 4.1 Continued from previous page

		ITS3P	GCATCGATGAAGAACGCAGC	(4)	
ITS2	Nuclear	ITS2P	GCTGCGTTCTTCATCGATGC	(4)	798 bp
		ITS5P	GGAAGGAGAAGTCGTAACAAG	(3)	

(1) (Olmstead and Sweere, 1994); (2) (Dong et al., 2015); (3) (Moller and Cronk, 1997); (4) (Moller and Cronk, 1997)

4.2.3 Phylogenetic analysis

Reference sequences of Miconieae available on NCBI (accession numbers listed in Table D.1 in Appendix D) were included in the analyses to evaluate the monophyly of *T. guianensis* and the species membership of the specimens collected. Optimal nucleotide substitution models were selected based on the Akaike Information Criterion (AIC) as implemented in **jModelTest2** (Darriba et al., 2012). Bayesian analyses for each locus were conducted using **MrBayes v3.2** (Ronquist et al., 2012), using a strict clock model and a GTR+G+I substitution model. Two runs were performed for each locus, with three heated chains and one cold chain of ten million states logging parameters every 1000 generations. Log files and effective sample size for all parameters were evaluated using **TRACER v.1.8.2** (Beast packages, Drummond et al. 2012), applying a burn-in of 10% of the total number of states. A 50 majority rule consensus tree was obtained using the sumt command in **MrBayes v3.2** (Ronquist et al., 2012).

To infer divergence times, separate ITS and ycf1b Bayesian phylogenies were calibrated using **BEAST v1.8.4** (Drummond et al., 2012) and based on the age estimations for Miconieae in Morley and Dick (2003). A normally distributed prior was placed on the tMRCA of all Miconieae with a mean of 65 million years (Ma) and a standard deviation of 10 Ma as the confidence intervals are not reported by Morley and Dick (2003). Confidence intervals were also selected to account for their estimates of the origin of Melastomataceae and Miconieae without imposing hard bounds (as the calibrations are not fossil records). The uncorrelated relaxed clock model was set to a prior exponential distribution with a mean of 18 Ma and a standard deviation of 0.33, based on the crown estimated age for Miconieae (Morley and Dick, 2003). The tree model prior used was the birth-death process (Gernhard, 2008). Two independent MCMC chains

of 300 million generations were run and parameters logged every 3000 generations. Additional runs without no data were carried out to confirm that priors were not biasing the posterior probabilities (Sanders and Lee, 2007). Log files and effective sample size for all parameters were evaluated using **TRACER v.1.8.2** (Beast packages, Drummond et al. (2012)). **LOGCOMBINER v.1.8.2** and **TREEANNOTATOR v.1.8.2** (Beast packages, Drummond et al. (2012)) were used to combine log and tree files from all the runs, applying a burn-in of 10% of the total number of states. All tree visualizations were done using Baltic (available at `https://github.com/blab/baltic`)

Ancestral areas of the ITS ultrametric tree were reconstructed using the Lagrange analysis implemented in **RASP** (Yu et al., 2015). Data from ITS suggests *Tococa* accessions are geographically structured. Based on this and on the Andean uplift times, three matrices of dispersal constraints were set as described in the *Methods* section in Chapter 2. Sequences from NCBI were excluded from the analysis. Finally, a Mantel test was performed to test the correlation between genetic and geographic distances and to differentiate isolation by distance (IBD) from vicariance promoted by the Andes. Under the vicariance model, genetic distances are expected to be large between populations on opposite sides of the Andes and small among populations in the same side, irrespective to geographic distances. Under the IBD model, genetic distances are expected to be positively correlated with geographic distances, with neighboring populations being genetically very similar no matter on which side of Andes they are located. In a scenario where the Andes uplift contributed to population divergence, geographic distances would not be good predictors of genetic distances. For the Mantel Test, pairwise sequence distance and pairwise geographic distance (measured as great-circle distance) for ITS and ycf1b were estimated using **Python v.2.7.** scripts.

4.2.4 DNA extraction and whole genome sequencing

The expected coverage is the depth at which each nucleotide is sequenced assuming that reads are distributed evenly across the genome (Sims et al., 2014). It depends on the size of the genome, the initial amount of DNA sequences and whether samples are pooled in the sequencing lane. The expected coverage is useful when designing genome sequencing experiments because allows for sequencing optimization: getting the best coverage given a number of samples and genome sizes per sequencing lane (Rabinowicz and Bennetzen, 2006). Thus, to calculate the expected coverage of the *T. guianensis* accessions to be sequenced it is necessary to calculate the size of *T. guianensis* genome. To confirm the genome size of *T. guianensis*, leaf material from three different individuals in the living collection at the Munich Botanic Garden was sent to Plant Cytometry Services (Kapel Avezaath, Netherlands) for a flow cytometry analysis. A sample of *Dissotia* (Melastomataceae) was used as a standard for the flow cytometry. Genome sizes estimated using flow cytometry and densitometry are estimations closer to real genome sizes than estimates derived from sequenced genomes, specially when genomes are large (Elliott and Gregory, 2015). The size of *T. guianensis* genome was calculated assuming diploidy. Knowing the number of genome copies of the standard sample is essential to correctly assign and report the haploid size of the genome in a standardized way (Doležel et al., 2007). Previous studies by Solt and Wurdack (1980); Almeda and Chuang (1992) and Almeda (1997) concluded that none of the cases of polyploidy in Melastomataceae occur in *Tococa* or *Dissotia*.

Quality and quantity of extracted DNA are other factors influencing choices on library preparation protocols and sequencing technologies. As mentioned above, extracting

DNA from *Tococa* is challenging. Hence that most DNA extractions from *T. guianensis* specimens yielded low concentrations of DNA, for Sanger sequencing and library preparations. The low DNA yields of my extractions were best suited to paired-end sequencing using short insert sizes, an approach that is also robust to mild DNA degradation.

Two runs of library preparation and genome sequencing were carried out. In the first, one *T.* cf. *guianensis* from Chocó and two from Putumayo (Figure 2.1 in Chapter 2) were selected for the first whole genome sequencing. The aim of choosing these is to sample the genetic variation within population and between geographically distant sites. Sample identification considered leaf, fruit and floral morphology, in addition to the results from the ITS, ndhF and ycf1b Sanger sequencing. DNA was extracted from two leaf fragments of 1x3 cm per sample, each processed in separate tubes and using the Qiagen Plant DNeasy kit reagents. The tissue was macerated using a TissueLyser II (Qiagen, Germany) for two minutes at 20 Hz. 800μL of AP1 buffer, 30μL Proteinase K (10mg/mL) and 30μL β-Mercaptoethanol were added to each tube before overnight incubation at 65°C. 1μL RiboShredder™RNase Blend (Illumina) was added to the mix and left incubating for 30 min at 37°C. 260μL P3 buffer was added to each tube before the ice incubation step. Contents from each tube were cleaned using different QIAshredder Mini spin columns, the resulting flow-through of each sample (not tube) were then pooled together in a DNeasy Mini spin column per sample. This modification aimed to increase the final DNA yield and reduce the amount of degraded DNA eluted. The rest of the extraction proceeded as described in the Sanger sequencing section. DNA integrity was assessed on a 2% agarose gel stained with 2.5μL ethidium bromide and quantified using the high sensitivity assay of QuBit DNA quantification system (Invitrogen).

DNA was fragmented using a Bioruptor Plus (Diagenode, Belgium) for five low power cycles of 30 sec on/90 sec off. Library preparation followed the TruSeq Nano LT kit for DNA samples (Illumina) for 550bp insert sizes for 200ng input DNA. During the last clean up step for amplified DNA, 37.5μL Sample Purification Beads were used instead of the 50μL indicated in the protocol. As larger amplicons bind first to the beads, using a lower bead concentration will prevent an excess of small fragments from binding and hence prevent these from reducing the quality of the final libraries. Input DNA and library quality were assessed on an Agilent 2100 Bioanalyzer (Agilent Technologies, United States) and quantified using the high sensitivity assay of QuBit DNA quantification system (Invitrogen). The resulting plant libraries were pooled together with the ant libraries from Chapter 3 to run them on one lane of the Illumina HiSeq 2500 platform in high-output mode.

A second effort to sequence the same specimens was carried out with the intention of increasing read coverage. Long read sequencing was not possible as the extracted DNA was too degraded to work with technologies like PacBio or Nanopore. For the second genome sequencing test, new material from the same specimens with an additional sample from Antioquia (AN in Figure 4.1) was treated as mentioned above and the following modifications. Library preparation followed the TruSeq Nano LT protocol for 350bp insert size using 100ng input DNA per sample. Library quality was assessed on the TapeStation and final sample concentration measured using the high sensitivity assay of QuBit DNA quantification system. Equimolar amounts of all specimens were pooled together and sequenced on one lane of the Illumina HiSeq 4000 platform in high-output mode.

4.2.5 Low coverage genome assembly

The quality of raw reads was assessed with **FastQC** (Andrews and others, 2010) and low quality read endings and PCR adapters were removed using **Trimmomatic v0.35** (Bolger et al., 2014). A preliminary assembly was done using a kmer size of 31bp in **Velvet v1.2.10** (Zerbino and Birney, 2008). To generate a set of nuclear contigs, it is necessary to filter out organelle reads (chloroplast and mitochondrion), and any non-plant contaminant DNA. The taxonomic affiliation of reads was checked with a Blastn search implemented on the command line tool **BLAST v2.6.0** (Camacho et al., 2009) against the curated non-redundant nucleotide (nt) NCBI database. Reads with a maximum e-value of $1e^{-25}$ were kept. Information about the read coverage on each contig was obtained by mapping all reads back into the contigs using **BWA-MEM v0.7.15** (Li and Durbin, 2009) and then converting the SAM file into BAM format with **Samtools v1.3.1** (Li et al., 2009). Taxonomic and coverage information were combined in 'blobplot' of the contigs using **Blobtools v0.9.19.5** (Kumar et al., 2013; Dominik R. Laetsch, 2017). A blobplot is a plot of all reads' coverage and GC content in the y and x axes respectively, whilst including color-coded labels of the subject taxa extracted from a blastn analyses of the contigs. This plot allows the detection of contaminant reads from non-target taxa (including, for example, plant endophytic fungi and bacteria, endophytophagous arthropods and nematodes) and helps further read filtering (Kumar et al., 2013; Dominik R. Laetsch, 2017).

All contaminant reads were removed if they matched to organisms other than Streptophytes (NCBI GI accessions under the Taxonomy ID:35493, last consulted April 2016). Removal from the assembly was done using the seqfilter option in **Blobtools**, after mapping the contaminant reads against the assembly with **BWA-MEM v0.7.15** (Li

and Durbin, 2009). **Blastn** reports different taxonomic matches for each contig if they occur. **Blobtools** scores **Blastn** matches based on the **Blastn** scores and the order of the hits. Then it reports the list of the hits with highest scores for each contig (Usually from the same taxonomic ID). In some cases, each contig can hit sequences with different taxonomic IDs. Those cases were checked by hand and if the score assigned to the Streptophyte hit is higher than the score assigned to the other taxonomic IDs by 200, the read will be considered as a plant contig, otherwise as a contaminant.

Once all contaminant, mtDNA and cpDNA reads were removed, kmer frequencies for the remaining pool of reads were counted using 19, 21, 41, 51, 71, 81 and 127 hash sizes with the intention to optimize the Velvet assembly (Figure D.2 in Appendix D). Smaller kmers than that are less effective at handling repeats, whilst using larger kmers increases the chances for those kmers to contain sequencing errors and the memory requirements (Bleidorn, 2017). If the memory needed to solve a Bruijn graph exceeds the memory capacity of the cluster the assembly usually crashes.

Different assembly strategies were tested using **Platanus v1.2.4** (Kajitani et al., 2014), **MaSuRCA v3.1.3** (Zimin et al., 2013), **dipSPAdes v3.6.2** (Safonova et al., 2015), **Velvet v1.2.10** (Zerbino and Birney, 2008) and **MetaSPAdes v3.10.1** (Nurk et al., 2017), from each the N50 and longest scaffold size were measured (Perl script provided by S. Kumar). Assemblers were chosen based on their power to deal with heterozygous genomes, repetitive regions, robustness under low coverage, and no need for a reference genome. Although the first assembly tests were carried out merging the reads of the first three specimens sequenced (PMFT244, PMFT466 and PMFT468), all subsequent reads were done individually, and different libraries were not combined. Assuming all *Tococa*-like specimens belong to the same species, reads from different samples can be

combined if they were sequenced in the same lane run. But combining different libraries from different runs increase the memory requirements and often crashed.

Genome assembly completeness was assessed using Benchmarking Universal Single-Copy Orthologs (**BUSCO v2**) (Simão et al., 2015). **BUSCO** identifies single-copy regions in the assembly by performing a **tBlastn** search against a database of conserved orthologous across subsets of organisms (*e.g.* plants, Hymenoptera, bacteria, etc.). Then, **BUSCO** reports the percentage of genes from the database that are complete, duplicated, fragmented or missing from the assembly (Simão et al., 2015). Conserved orthologous genes are expected to be complete as they are likely under selection; fragmented or missing genes in the assembly are an indication that the contigs are not correctly assembled. Single gene annotation is performed by **BUSCO** using the **Augustus** protein prediction algorithms. **BUSCO** was run on the *Tococa*-like assemblies using the embryophyte database consisting of 1,440 conserved orthologue genes and using *Arabidopsis* as the default species parameters for **Augustus**.

Finally, the assembly with the best completeness record was used as a reference to assist the scaffolding of the remaining assemblies using **AlingGraph** (Bao et al., 2014) and **Scaffold Builder** (Silva et al., 2013). It is better to use a reference genome at the scaffolding step and not at the assembly step because the expanding process (*i.e.* the merging of contigs if they overlap or are close according to the reference genome) might introduce erroneous gaps on the final assembly. The assembly step infers the position of reads and connect them based on the reference genome. If the reference has TE, inversions or it is incompletely assembled, the resulting assembly will have large contigs of few reads connected by hundreds of Ns (uncalled or ambiguous bases), not to mention that the position of those reads can be misplaced. In contrast, assisted scaffolding will

only help to resolve the position of reads based on the reference and with respect to the rest of the reads, without incorporating alien information into the final assembly. Finally, using transcriptomes as references would have only dealt with coding regions but would have left most of the assembly unresolved. The recommended procedure to merge assemblies generated from different libraries is by following the Genome Analysis Toolkit (**GATK**, Van der Auwera et al. 2002) best practice pipeline which can take months (but will be done in the future). For this reason, only the best assemblies of each sample were chosen for the following analyses.

4.2.6 Phylogenomic analyses

To explore the evolutionary relationships between the genomes selected and to screen candidate markers, all the **BUSCO** single-copy genes shared across the assemblies were aligned with **Muscle v3.8.31** (Edgar, 2004) and Maximum-Likelihood estimations of the trees and branch lengths were conducted using **RAxML v8.2.9** (Stamatakis, 2014) and 500 bootstrap replicates. Best bipartition trees were summarized and a species tree from the gene trees estimated using **ASTRAL v4.10.12** (Mirarab et al., 2014). **ASTRAL** uses the multi-species coalescent model to estimate a species tree accounting for the number of genes supporting each branch. More details about **ASTRAL** and species tree estimations are found in Chapter 3. Nucleotide diversity, the number of segregating sites and the Watterson's Θ from the alignments were estimated using the Python's Dendropy package (Sukumaran and Holder, 2010). These statistics are indicators of nucleotide substitutions across sequences and the diversity in a population, controlling for alignment length, thus they are indicators of the phylogenetic information on the sequences.

The appropriate criteria for selection of candidate markers for phylogenetic studies depend on the goal. For instance, rapidly evolving genes (with high nucleotide diversity and many segregating sites) have greater potential to resolve relationships within populations, whilst slow evolving genes with fewer segregating sites will be more useful to resolve relationships between species and higher taxa (genera, tribes, etc.). Thus, rapidly evolving genes provide the best hope for well-supported resolution of relationships between *T. guianensis* populations than slower genes. Candidate markers that will perform well under different scenarios were selected based on the number of segregating sites of each single-copy **BUSCO** gene alignments of the four *T. guianensis* genomes: two from Putumayo, one from Chocó and one from Antioquia. The quartiles of the distribution of segregating sites across alignments were used to classify the **BUSCO** genes into four categories: low diversity (first quartile), medium diversity (second quartile), high diversity (third quartile) and super-high diversity (fourth quartile). From each category, 50 random genes were selected and used to estimate gene and species trees using **BEAST v.1.8.4** (Drummond et al., 2012). Two MCMC chains of 30 million generations were run assuming a GTR+G model of nucleotide substitutions and a strict clock and coalescence with constant population size models (Heled and Drummond, 2008). A hyperprior with mean 0.0003 and standard deviation 0.01 was set for the evolutionary rate, such that the distribution includes the evolutionary rate estimates used in Morley and Dick (2003) and those obtained in this study. To identify which quartile provided the best resolution, a Maximum Clade Credibility Species (MCCS) tree was estimated from the majority consensus species tree of each majority consensus gene trees using **TREEANNOTATOR v.1.8.2** (Beast packages, Drummond et al. 2012). Branch posterior probabilities were obtained from the MCCS tree and results for each category were visualized using **DensiTtree v.2** (Bouckaert and

Heled, 2014).

4.3 Results

Tococa cf. *guianensis* plants were collected in most sites planned originally, in most cases inhabited by ants and usually growing close to water sources. However, no individuals were found in the areas between the Central and West Cordilleras (2, 3, and 4 in Figure 2.1 in Chapter 2), despite the presence of forested patches like those where *T. guianensis* is usually found. Finding the plant was particularly difficult in areas of the Santander region (Barrancabermeja and Cimitarra in Figure 2.1 in Chapter 2) due to extreme habitat disruption and to removal because people have a general dislike of the plant for bearing ants. Habitat fragmentation negatively affects communities of ant-plant mutualists, lowering their population sizes and reducing the long-term persistence of such communities (Bruna et al., 2005). The number of specimens collected in each area is listed in Table B.2 in Appendix B. Individuals were usually found in small patches of five to ten plants of heights between 1-2 m. Taller plants are often pruned by people and rarely found unless they are growing in a conserved forest. Domatia are found in young plants after the first pair of leaves has developed, and ants can be found inhabiting these even when the plant has not branched and bears only a few leaves. Across all collecting sites and sometimes within the same area (*e.g.* Amalfi, Puerto Nariño, and Villagarzon in Figure 2.1 in Chapter 2), I found large variation in the level of anisophylly (difference in size of a pair of opposite leaves), leaf size, trichome density, and vein color -all traits used to identify *T. guianensis*. That variation introduced uncertainties in the morphological identification of the specimens, as one specimen sometimes fits the description of more than one *Tococa* species. For this

reason, I refer to the *T. guianensis* specimens in this project as *Tococa*-like specimens, although all plants included in the analyses fit the description of *T. guianensis*.

4.3.1 Phylogenetic inference

The ITS region, compared to ycf1b and ndhF, has more segregating sites and a higher population genetic diversity. After trimming low quality segments, sequences resulted in fragments of between 767-778 bp with 270 segregating sites and Watterson's Θ= 47.18. The ITS phylogeny shows a three clade polytomy (Figure 4.1) that is concordant with the results presented by Goldenberg et al. (2008). The first clade is *Miconia cinerea* section Miconia (Goldenberg et al., 2013) represented here by a single species. The second clade (clade **A**) is a highly-supported (higher than 0.9 posterior probability) clade of mostly Amazonian *Tococa*-like specimens, including three specimens collected to the west of the Eastern cordillera, *T. caquetana*, *Maieta*, and one *Miconia*, forming the *Clidemia* grade within section *Miconia* (as Goldenberg et al. 2013 refers to the group). A further poorly supported clade (clade *B*, with higher than 0.7 posterior probability) includes four subclades in a polytomy with *Mi. ferruginea* section Chaenopleura as a singleton. The next subclade corresponds to the *Conostegia* (**C**) *sensu lato* clade and includes *T. spadiciflora*, *Conostegia* and a *Tococa*-like specimen. Subclade **D** corresponding to the *Mecranium*, *Anaectocalyx* and allies clade (Goldenberg et al., 2013) includes *T. platyphylla* and the sister species *T. broadwayi* and *T. perclara*. The last subclade **E** includes mostly *Tococa* NCBI references and *Tococa*-like specimens. *T. raggiana*, *T. bolivariensis*, *T. rotundifolia*, *T. macrophysca* and *T. nitens* are placed within *Tococa sensu stricto* and branch earlier from the rest of the *Tococa* group. *T. macrosperma*, *T. guianensis* are in a polytomy with other *Tococa*-like specimens from Meta

and Antioquia, one clade with poor support (less than 0.7 posterior probability) including specimens from locations on both sides of the Andes, and two groups with support higher than 0.7 posterior probability and restricted to either Meta or Amazonas. This last group of Amazonian specimens also includes *T. caudata*, *T. capitata*, *T. coronata* and *T. subciliata*. Except for *T. gonoptera*, the topology recovered with ITS follows the same topologies previously recovered using nuclear and chloroplast DNA (Michelangeli et al., 2004, 2008). The position of *T. gonoptera* is resolved within a separate group with the rest of *Tococa*-like specimens, but its position in previous phylogenies had low support. Moreover, it falls within the same group as *T. discolor*, and it is important to note that *T. discolor* was assigned as a synonym of *T. guianensis* (Michelangeli 2005, the name *T. discolor* is retained here to be consistent with NCBI records). Reference sequences of *T. guianensis* and *T. discolor* belong to different groups, also suggesting that *T. guianensis* is paraphyletic. This last group is also composed of a grade of *Tococa*-like specimens from east of the Andes that are distinct from a clade of specimens predominantly found to the west of the Eastern cordillera with only two exceptions.

The sequencing of ycf1b resulted in fragments of 776 bp, 102 segregating sites and Watterson's Θ= 17.44 from 193 specimens representing most collecting sites. Population diversity and number of segregating sites are lower than ITS, but higher than ndhF. The ycf1b phylogeny (Figure 4.2) is not as well resolved as the ITS phylogeny and shows a polytomy of most specimens, with four clades composed of four or more specimens showing some geographic structure (**A** to **D**). No ycf1b reference sequences of the tribe Miconieae were available in NCBI at the time the phylogeny was estimated (July 2016) and only two *Miconia* whole chloroplast genomes were recently made available. Results from ycf1b consistently recover *Henriettella* as the sister to the other Miconieae species included in this analysis. Within the rest of Miconieae, *Maieta* appears to be

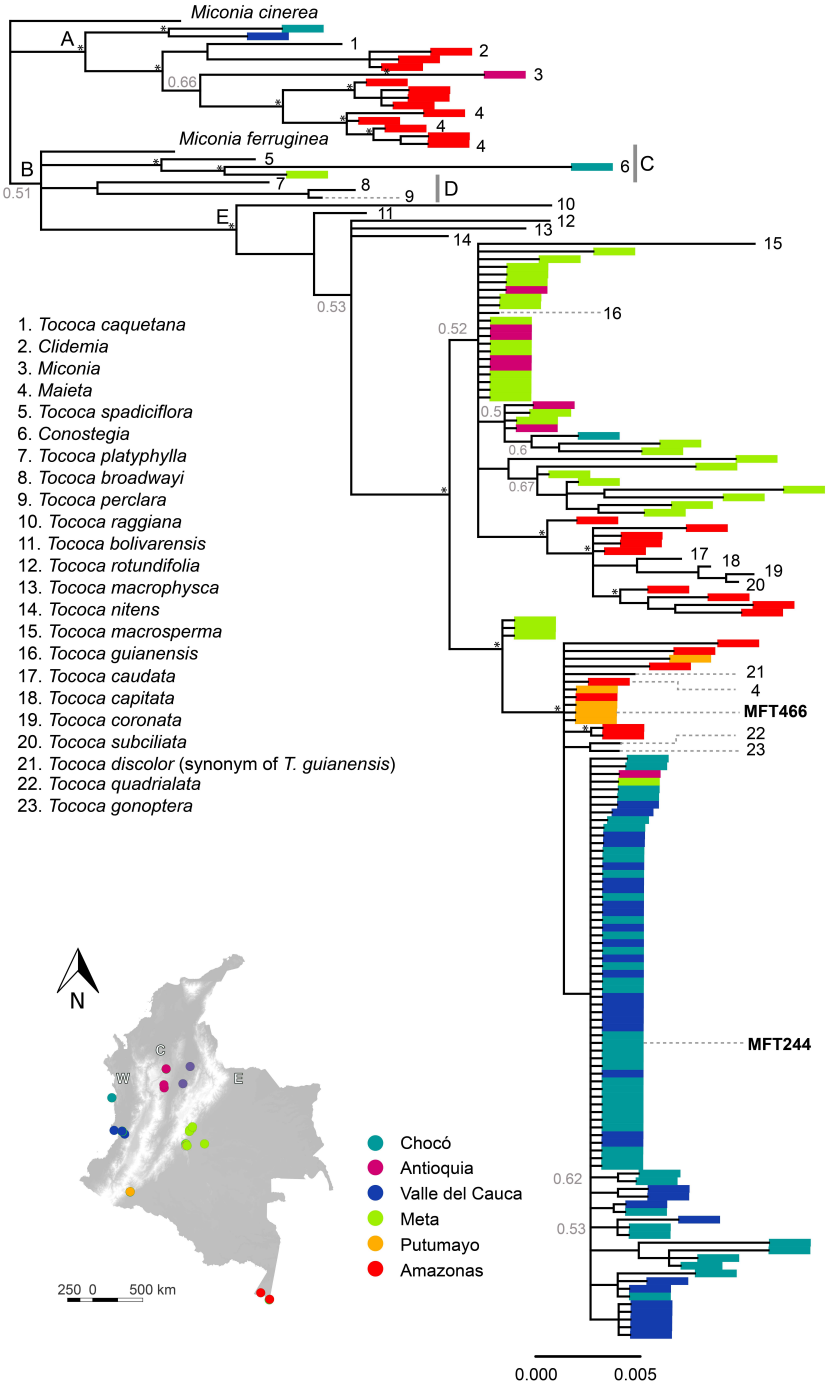


Figure 4.1: Maximum clade credibility ITS phylogeny of *Tococa*-like specimens. Branches with posterior probabilities lower than 0.4 are collapsed. Only posterior probabilities between 0.4 and 0.7 are shown, those not shown are higher than 0.7. Posterior probabilities higher than 0.9 are marked with an asterisk. Capital letters indicate clades as mentioned in the text, and tips are color coded according to the regions showed in the map of Colombia. Numbers indicate the position of the NCBI reference sequences included in the analysis.

paraphyletic with one specimen grouped in a clade with *Tococa*-like specimens from Amazonas and Putumayo (**C**), to the east of the mountain chain as in the ITS tree. The other *Maieta* is in a polytomy with numerous other *Tococa*-like specimens. Only one specimen of *Miconia* and *Clidemia* were sequenced and therefore the monophyly of these genera cannot be assessed. The other three clades include *Tococa*-like specimens from east of the Andes (**A** and **B**) and by *Tococa*-like specimens from west of the Eastern Cordillera (**D**). Unlike ITS, Clade **D** of Western *Tococa* does not fall within a clade of Eastern *Tococa*, but its relationships to other clades is not supported in the ycf1b tree. The other clades with less than four specimens show geographic structure (specimens are either from east or west of the Eastern Cordillera) except for one clade where one specimen from Antioquia and another from Meta cluster together.

Sequencing of ndhF was attempted by using internal primers but for most samples only the first section was successfully amplified, resulting in a fragment of 251 bp, Watterson's Θ= 7.09 and only 30 segregating sites, much lower in comparison to ITS and ycf1b. The paraphyly for all genera resulting in the ndhF phylogeny (Figure 4.3) and confirmed in the ITS and ycf1b phylogenies can be attributed to real paraphyly or to lack of resolution in the marker, likely due to short sequence length, failed amplification and lack of variation. The biggest clade forms a polytomy with singletons and four smaller clades each of two sequences. Within this, there is a well-supported clade of only *Tococa*-like specimens from Amazonas. The outgroups to those are *Clidemia dentata* and *C. rubra*, both part of a polytomy. Unlike ITS and ycf1b, ndhF shows little geographic structure except for the well resolved clade with Amazonian samples.

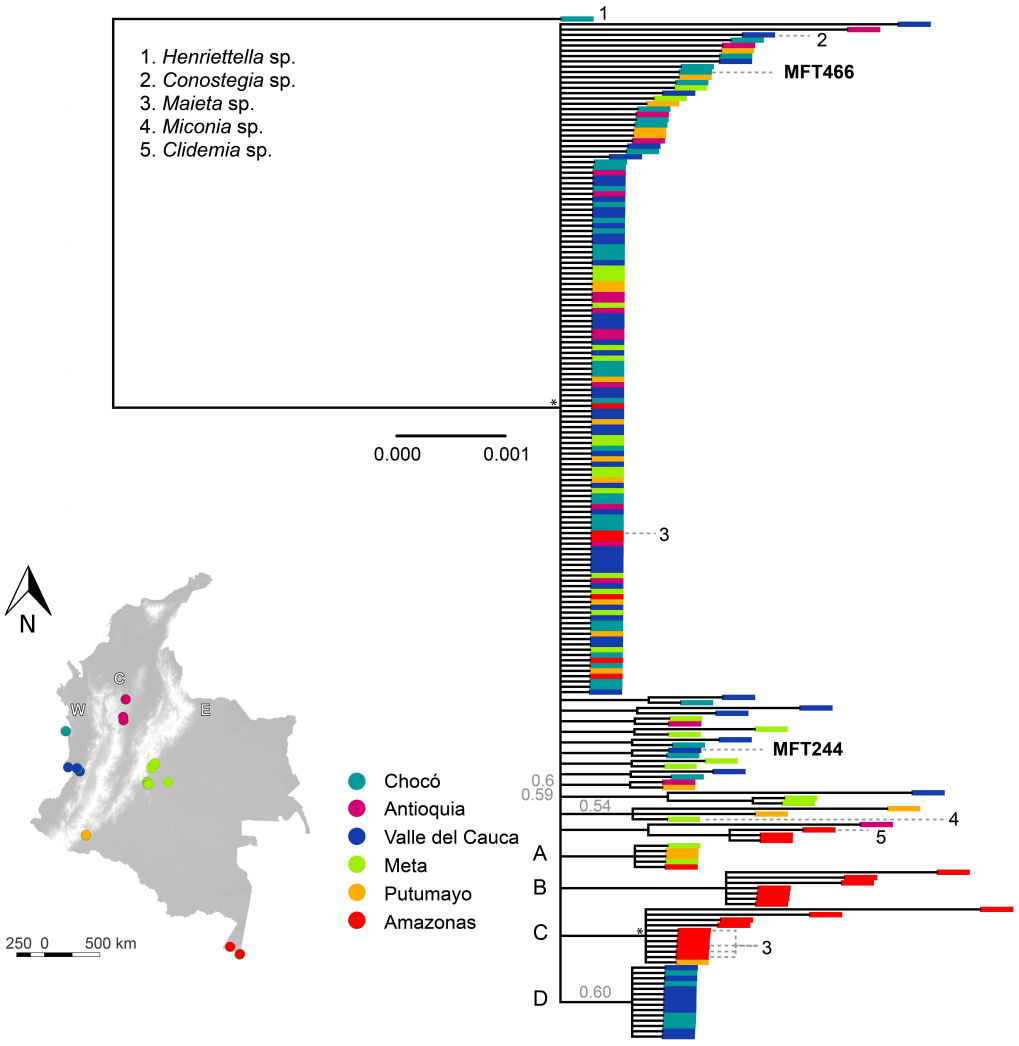


Figure 4.2: Maximum clade credibility ycf1b phylogeny of *Tococa*-like specimens. Branches with posterior probabilities lower than 0.4 are collapsed. Only posterior probabilities between 0.4 and 0.7 are shown, those not shown are higher than 0.7. Posterior probabilities higher than 0.9 are marked with an asterisk. Capital letters indicate clades as mentioned in the text, and tips are color coded according to the regions showed in the map of Colombia. Numbers indicate the position of the NCBI reference sequences included in the analysis.

4.3.2 Tree calibrations and geographic reconstructions

The ITS calibrated phylogeny (Figure 4.5) shows significant geographic structure: early divergent clades are distributed east to the Andes and the most recently diverged clade corresponds to two lineages, one from east to the Andes and the other from west to the Eastern Cordillera. Moreover, such structure coincides in time with major changes in

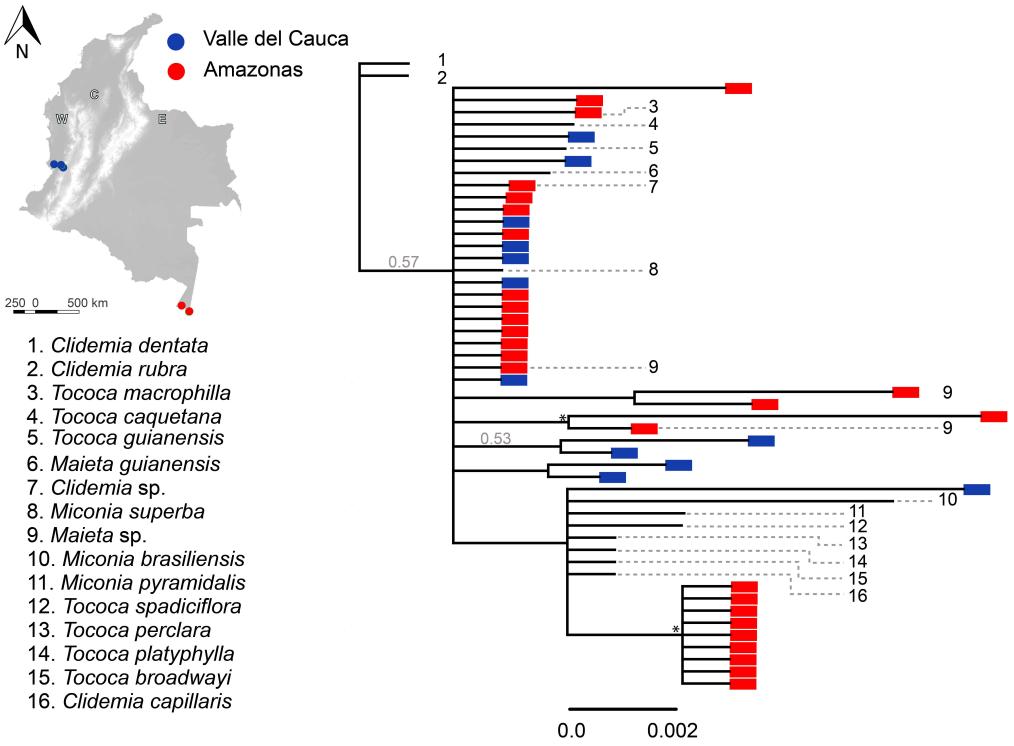


Figure 4.3: Maximum clade credibility ndhF phylogeny of *Tococa*-like specimens. Branches with posterior probabilities lower than 0.4 are collapsed. Only posterior probabilities between 0.4 and 0.7 are shown, those not shown are higher than 0.7. Posterior probabilities higher than 0.9 are marked with an asterisk. Capital letters indicate clades as mentioned in the text, and tips are color coded according to the regions showed in the map of Colombia. Numbers indicate the position of the NCBI reference sequences included in the analysis.

the Andes. In Figure 4.5, estimates for the tMRCA of all Miconieae (clades **A** and **E**) are around 50 Mya (95% CI = 31-66 Mya), with confidence intervals consistent with the ages reported by Morley and Dick (2003), suggesting that Miconieae was already established in the Neotropics 55 Mya during the Eocene. Later splits occurred within **E** separating a lineage of other *Tococa* species from two eastern lineages of *Tococa*-like specimens at 33 Mya (95% CI = 17-49 Mya), the largest of the lineages subsequently divided into the eastern lineages and one western lineage diverging within them at 16 Mya (95% CI = 7.9-26 Mya, marked with a star in Figure 4.5). In few exceptions specimens collected west to the Eastern Cordillera fall within eastern lineages, as is the case of Antioquia specimens falling among specimens from Meta (marked with **C**).

Reconstruction of ancestral areas for the *Tococa*-like clades places the Most Recent Common Ancestor (MRCA) to all *Tococa* in Colombia in an area east to the Andes that encompasses Amazonas, Meta and Putumayo (B, D, and E in Figure 4.4), possibly including either Valle del Cauca or Antioquia (A and E respectively). The same resulted for the MRCA to the mostly Amazonian clade **A** and the MRCA to all Colombian samples within clade **E**. Within clade **E**, the area of the MRCA for the Amazon and Meta lineages, sister to clade **C**, is likely to be either Amazonas or Meta, both east to the Andes. MRCA to clade **C** is estimated to be Meta, suggesting several independent migrations to Antioquia. Similarly, the ancestral area of the MRCA to the Andean lineage (marked with a star in 4.4) is the same as the ancestral area to all Colombian *Tococa*, suggesting that the lineage could have expanded east to west to then be split by the Andean uplift. Subsequently, the western lineage expanded to Chocó, with a migration to Antioquia and Meta.

A Mantel test was applied to test the correlation between genetic and geographic distances. Genetic distances were calculated as pairwise sequence identity and geographic distances were calculated as great-circle distances between specimens' coordinates. Pairwise genetic distances are higher for ITS than for ycf1b (Figure 4.6), and according to the test, geographic distances explain more of the variance in genetic distances between ITS sequences (13.8%) than between ycf1b sequences (1.66%). Although overall, the proportion of variance in genetic distances explained by geographic distances is low. Genetic pairwise differences (lower matrix lft panel, Figure 4.6) within neighboring populations are expected to be low; however, distances between distant and neighboring populations are similar, except for the Amazonas sequences. But although Amazonas (AM in Figure 4.6) is the most geographically distant population, genetic differences within Amazonas and between Amazonas and other populations are low

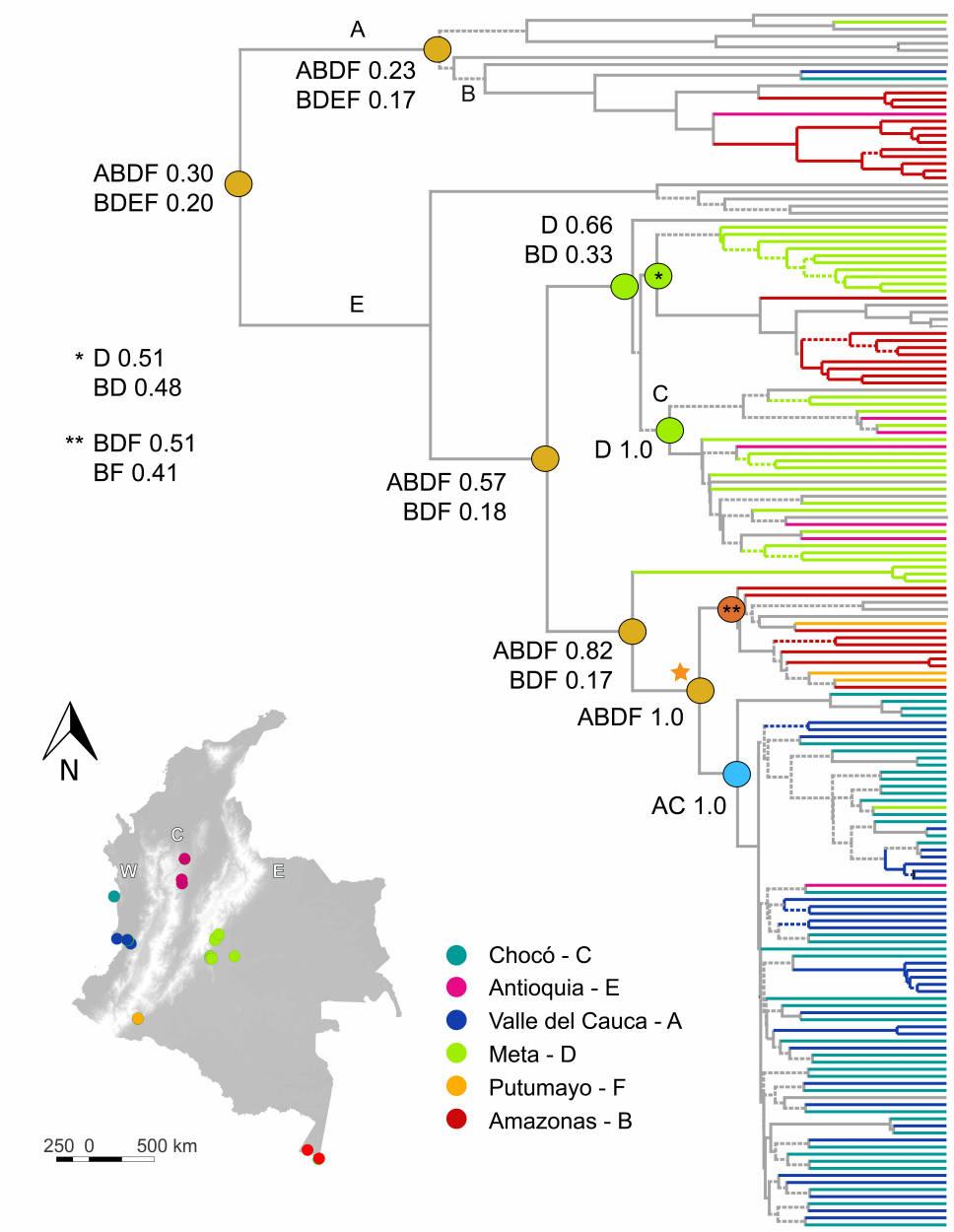


Figure 4.4: Calibrated majority consensus tree obtained from the large dataset of ITS2 sequences and the reconstruction of ancestral areas for the nodes of *Tococa*-associated *Azteca*. The two (or more) most probable ancestral areas and their probabilities are shown for the nodes of interest. The star indicates the split between Western and Eastern *Azteca*. The map shows the collecting sites and the color code for the specimens. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

in half the cases (right panels in Figure 4.6). Results from the Mantel test show a positive correlation between geographic and genetic distances for both, ITS (r= 0.372, P= 0.0001, Z= 8.36) and ycf1b (r= 0.129, P= 0.02, Z= 2.428), but the slope for the correlations is shallow. These patterns suggest that the Andes acts, and likely acted as a barrier to dispersal during the diversification of the *Tococa*-like lineages and that geographic distances are not key determiners of genetic divergence.

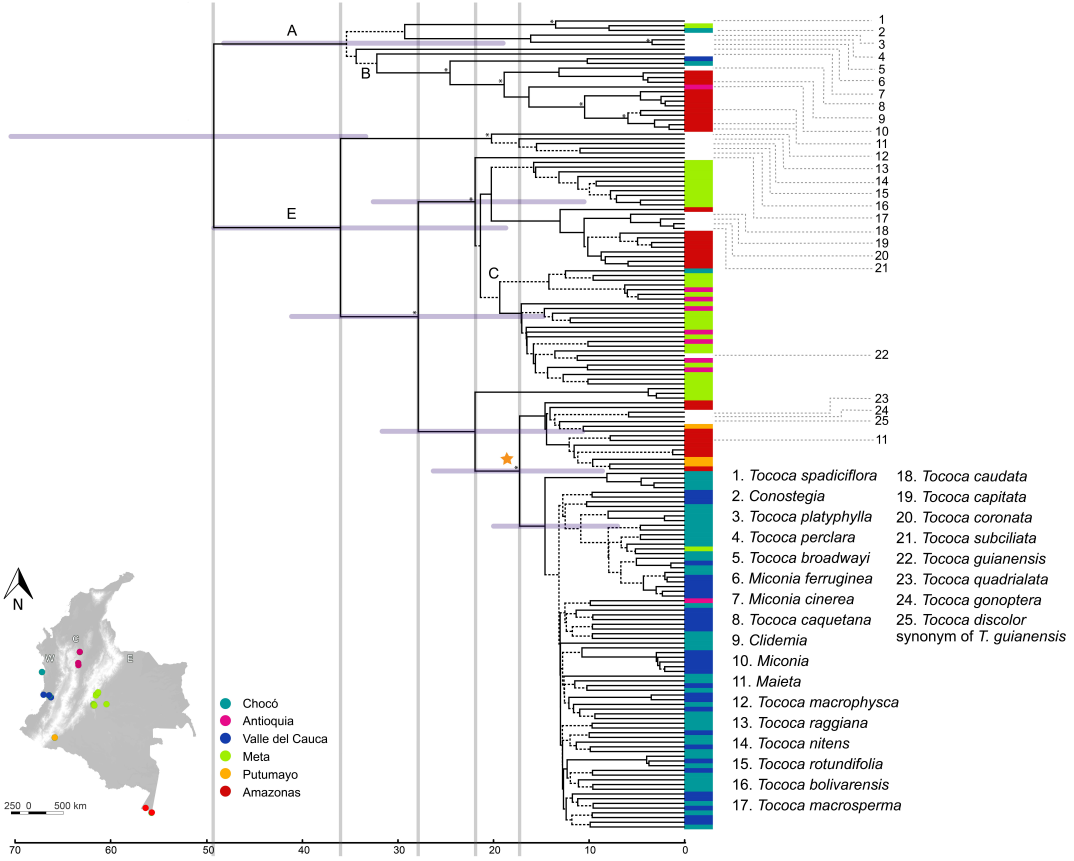


Figure 4.5: Maximum clade credibility ITS calibrated phylogeny of *Tococa*-like specimens. Branches with posterior probabilities lower than 0.4 are collapsed. Only posterior probabilities between 0.4 and 0.7 are shown, those not shown are higher than 0.7. Posterior probabilities higher than 0.9 are marked with an asterisk. Capital letters indicate clades as mentioned in the text, and tips are color coded according to the regions showed in the map of Colombia. Numbers indicate the position of the NCBI reference sequences included in the analysis.

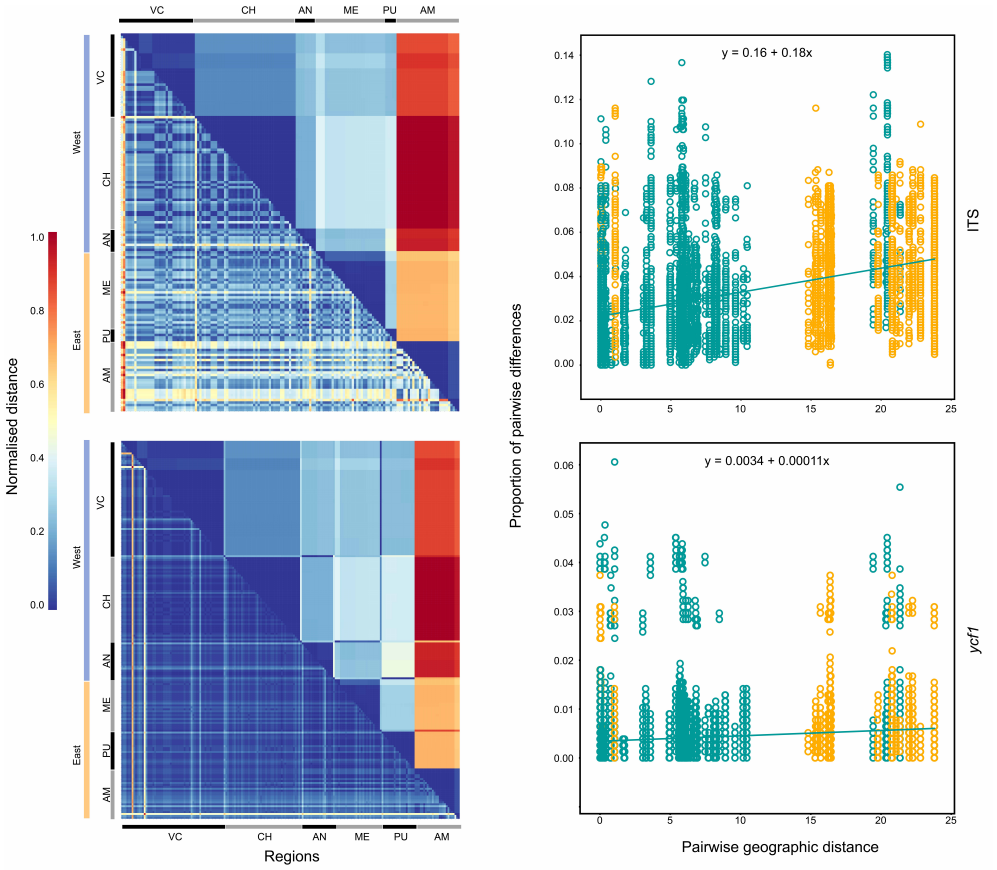


Figure 4.6: Left panels: Heat maps showing pairwise genetic and geographic distances between ITS and ycf1b sequences. Distances have been normalized to make them comparable. In each panel, values above the diagonal show the geographic distances calculated as great-circle distances to correct for the Earth's circumference. Lower matrices show genetic distances calculated as the proportion of differences between sequences. Abbreviations for the collecting places correspond to those in Figure 4.1. Right panels: Genetic distances plotted against geographic distances. Orange represents pairwise comparisons involving one or both sequences from the Amazon, green represents otherwise. The fitted line corresponds to the linear regression for the correlation between genetic and geographic distances.

4.3.3 *De novo* genome assembly

T. guianensis genome size was estimated to be 0.69 pg/2c (an average of 0.346 pg/c across samples), corresponding to a C-value of 339 Mb (For the flow cytometry see Figure D.1 in Appendix D). Because the phylogenetic analysis of ITS shows that *T. guianensis* is paraphyletic (or not a single species at all), the selection of samples for the

de novo genome sequencing was based on geography more than molecular identification, to ensure that the sampling of genomes includes geographically close and distant samples. Thus, four fertile specimens identified as *T. guianensis* were selected, two from Putumayo south-east of the Andes, one from Choco west of the Andes, and one from Antioquia between the Western and Central Andean Cordilleras (Figure 4.11 and Table 4.2). Two libraries from each specimen were sequenced (as explained in Methods and except for the MFT584 sample) each assembly differentiated by a **P** or a **T** before the specimen code (Table 4.2). Blast results from comparing all contigs against the curated nucleotide NCBI database identified an average of 606019 as contaminants. Proteobacteria and Ascomycota were the most common contaminants, the rest being other bacteria, fungi, and a negligible number of reads matching Chordata (Figure 4.7). Blobplots of the contig coverage identified around 343 plant contigs with higher coverage likely corresponding to plastid DNA that were removed to facilitate the nuclear genome assemblies. After removing up to 38.6% of contaminant reads, Illumina adaptors and low-quality reads, between 70 and 150 million reads remained (Table 4.2). Kmer counts after cleaning reads suggest the presence of repetitive sequences in all samples as a high number of frequently seen kmers, regardless of the kmer size (Figure D.2 in Chapter D). However, there is a tendency to a higher number of less observed kmers as the kmer size increases. Longer kmers provide better resolution for repetitive sequences and increase the length of output contigs, but memory requirements for genome assembly increase as well.

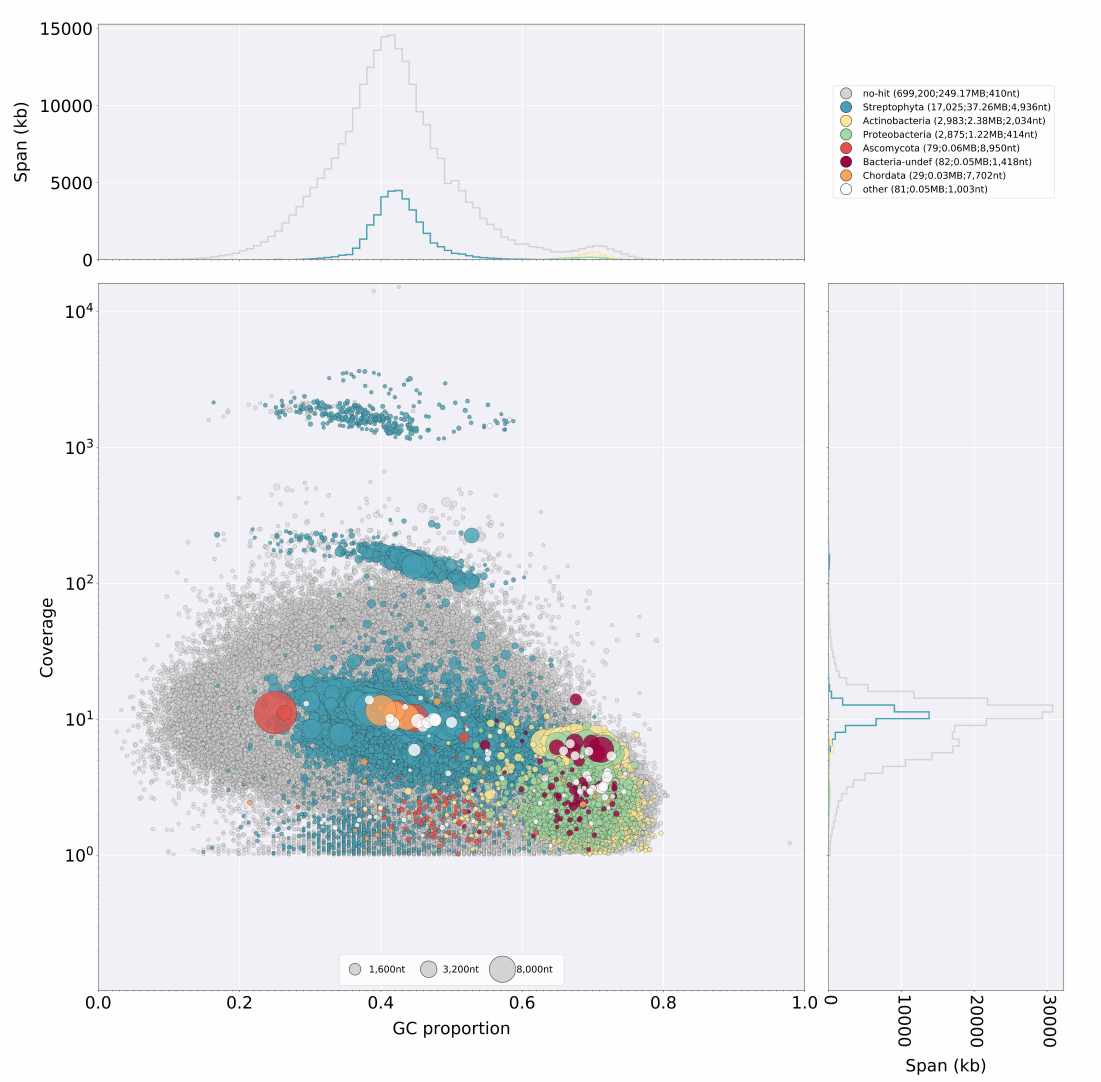


Figure 4.7: Example of a Blobplot generated for the PMFT244 *Tococa*-like assembly. Each blob represents a contig with a size proportional to the size in bases and a color corresponding to the taxonomic ID of the contig (extracted from Blast results). Three clouds of plant contigs at different coverage ranges can be identified: the bigger cloud corresponding to the nuclear DNA and the other two corresponding to either mitochondrial or chloroplast DNA.

Table 4.2: Number of Illumina pair end reads (in million reads), trimmed pair end reads, pair end reads after removing contaminants and plastid DNA, and %GC content for two library preparations and four *Tococa*-like specimens

Library	Specimen	Region	Total reads	Trimmed reads	After removal	Coverage*	%GC
P	MFT244	Choco (CH)	86.93	84.17	78.04	31.22	40
	MFT466	Putumayo (PU)	81.31	79.3	63.04	25.22	40
	MFT468	Putumayo (PU)	88.16	85.72	66.39	26.55	42
T	MFT244	Choco (CH)	175.27	165.65	118.45	47.38	42
	MFT466	Putumayo (PU)	173.41	164.44	106.47	42.59	42
	MFT468	Putumayo (PU)	160.31	151.99	110.13	44.05	42
	MFT584	Antioquia (AN)	161.36	151.64	141.8	56.72	41

*Approximate coverage assuming all reads are 120 bases long (the distribution ranges from 20-150 bases)

Different strategies were applied to assemble the reads from each run (Table 4.3). The first strategy attempted to generate a general reference genome for *T. guianensis* by combining the reads from the three specimens sequenced first (P library) using **Platanus** (Kajitani et al., 2014). The resulting assemblies were highly fragmented, only 18 out of 1440 genes were complete and the N50 statistic was very low independently from the parameters used for the assemblage. Thus, reads from each specimen and library were separately assembled as merged assemblies are of low quality. However, **Platanus** also failed to produce long contigs even when reads from different individuals were assembled separately, except for PMFT468. No attempt was made to assemble reads from the second run as expectations of a good assembly were very low. **Velvet** (Zerbino and Birney, 2008) performed better than **Platanus** only when kmer size= 71 was used, otherwise the assemblage process crashed. Successful **Velvet** assemblies did not improve much after the additional scaffolding step using **Scaffold Builder** and with the most complete **Velvet** assembly as a reference. Neither the contig length nor the assembly completeness increased substantially. While scaffolds joined with **Scaffold Builder** were not significantly longer, **Alignngraph** ran for more than three weeks with no significant output and had to be terminated. A final and more successful attempt to assemble the genomes used the **MetaSPAdes** (Nurk et al., 2017) package included in the **SPAdes v.3.11** assembler (Nurk et al., 2013). This resulted in higher N50 values than the **Velvet** assemblies in all but one sample. Other assemblers such as **Masurca** (Zimin et al., 2013) and **dipSPAdes** (Safonova et al., 2015) designed to deal with diploid, large, and repetitive genomes completely crashed (data not shown).

Completeness, or the percentage of complete single-copy genes in an assembly, is 81% for the PMFT244 **Velvet** assembly, but close to 50% or lower for the rest of assemblies (Figure 4.8). A high percentage of single-copy genes is reported as missing likely due

to fragmentation and up to 6% of the 1440 orthologues from the **BUSCO** database is duplicated. On the other hand, completeness percentages of the **MetaSPAdes** assemblies are higher than 70% in all but one case; however, the percentage of duplications increased slightly compared to the **Velvet** assemblies. Whether if those duplications are artefacts of the assembly or represent gene duplications was not tested, but the similar percentage of duplications in the assemblies across different assemblers suggests that they can represent actual duplications.

Table 4.3: N50 values for different strategies and kmer sizes (K) of the *Tococa de novo* genome assemblies. Numbers within brackets indicate the percentage of complete single-copy genes reported by BUSCO. Read used for all assemblages were filtered and contaminants and pDNA removed, but assembler options for clean reads were only specified when indicated.

	Platanus				**Velvet**					**MetaSPAdes**
Sample	**Merged**	**Merged clean***	**Merged, clean, Mdbc***	**Individual**	**K 31**	**K 41**	**K 41 clean****	**K 71**	**K 71 scaffold**	**All Kmers**
PMFT244				651	crash	crash	crash	14823 (81)	Reference	9182 (81.1)
PMFT466	350 (18)	347	312	519 (19)	1648	2405	2405	3782 (53.2)	8709 (54)	3717 (74.9)
PMFT468				2072 (65)	1128	1363	1363	3219 (49.2)	7943 (48.7)	3489 (71.2)
TMFT244	-	-	-	-	crash	crash	crash	2727 (59.2)	4807 (62.8)	5654 (76.5)
TMFT466	-	-	-	-	1166	crash	1219	2157 (36)	46686 (38.8)	2926 (63.5)
TMFT468	-	-	-	-	1063	crash	1424	1356 (28.2)	12473 (32.7)	2357 (59.8)
TMFT584	-	-	-	-	crash	crash	crash	2768 (56.6)	6038 (58.3)	4550 (77)

Table 4.3 continued from previous page

*Merged reads from the indicated specimens; clean means the command -very_clean=yes was specified;

Maximum difference for bubble crush (Mdbc) for high heterozygosity -u=0.2

** Specifying the option -very_clean=yes

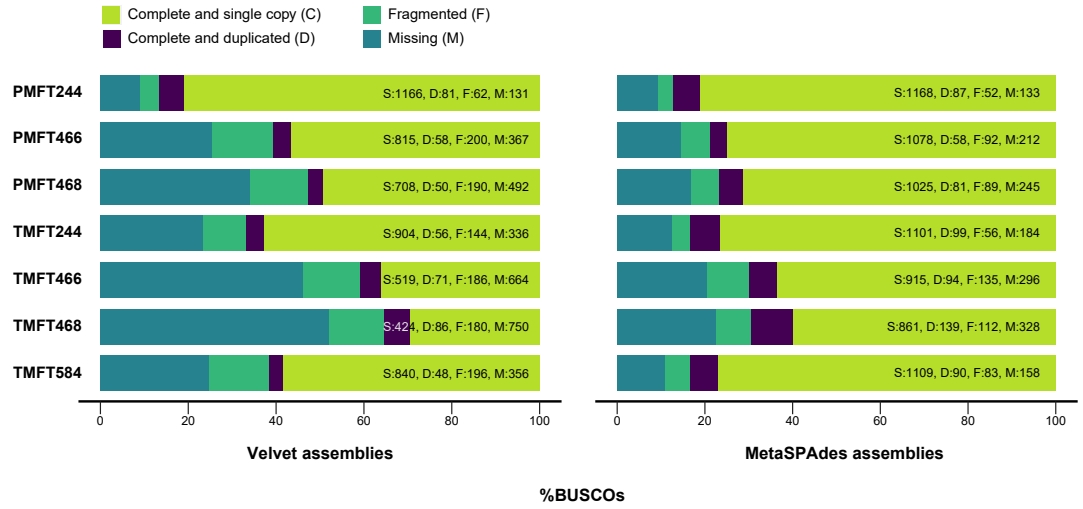


Figure 4.8: Percentage of completeness as the proportion of complete, duplicated fragmented and missing genes out of 1440 genes in the embryophyte **BUSCO** database. Bars on the left correspond to the completeness of **Velvet** assemblies and bars on the right to **MetaSPAdes** assemblies.

4.3.4 Phylogenomic analyses

The best assembly for each specimen was selected based on the N50 statistic and the **BUSCO** completeness, and only those were used for the subsequent analyses. Of the single-copy **BUSCO** genes predicted from the assemblies, 759 are shared across the four of them. The length of their alignments varies from 213 to 16693 bases long and the distribution of the nucleotide diversity, number of segregating sites and Watterson's Θ across the 759 alignments have means of 0.0308 ($\pm$0.0506), 70.656 ($\pm$115.571) and 38.540 ($\pm$63.039) respectively (Figure 4.9 and Table D.2 in D). Moreover, segregating sites, nucleotide diversity and the Watterson's Θ do not change in relation to the alignment length (Figure 4.10).

The **ASTRAL** unrooted species tree was estimated using the 759 single-copy **BUSCO** genes and shows the two Putumayo *Tococa* more closely related to each other than to *Tococa* from either Antioquia or Choco, with a branch local posterior probability of

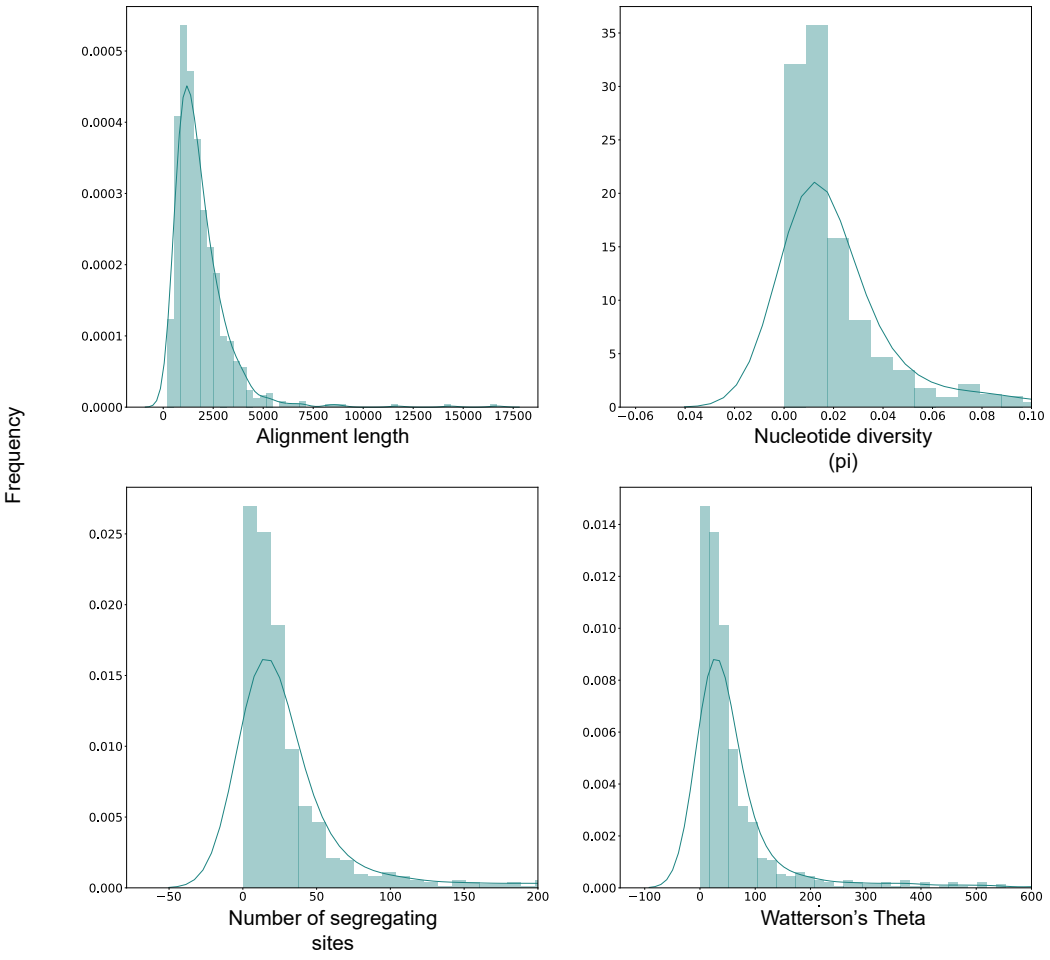


Figure 4.9: Frequency distributions of the length, nucleotide diversity, numbers of segregating sites and Watterson's Θ for the 759 alignments of the single-copy **BUSCO** genes shared across the four *Tococa* assemblies.

one (local posterior probability is described in more depth in Chapter 3). ASTRAL calculates branch support based on the frequency of quadripartitions on a tree and does not estimate branch support for terminal branches (Figure 4.11).

Single-copy **BUSCO** genes were classified into quartiles based on the number of segregating sites. Low diversity genes grouped in the first quartile have between 0 and 17 segregating sites, medium diversity genes have between 17 and 35, high diversity genes have between 35 and 66, and super-high diversity genes have between 66 and 865

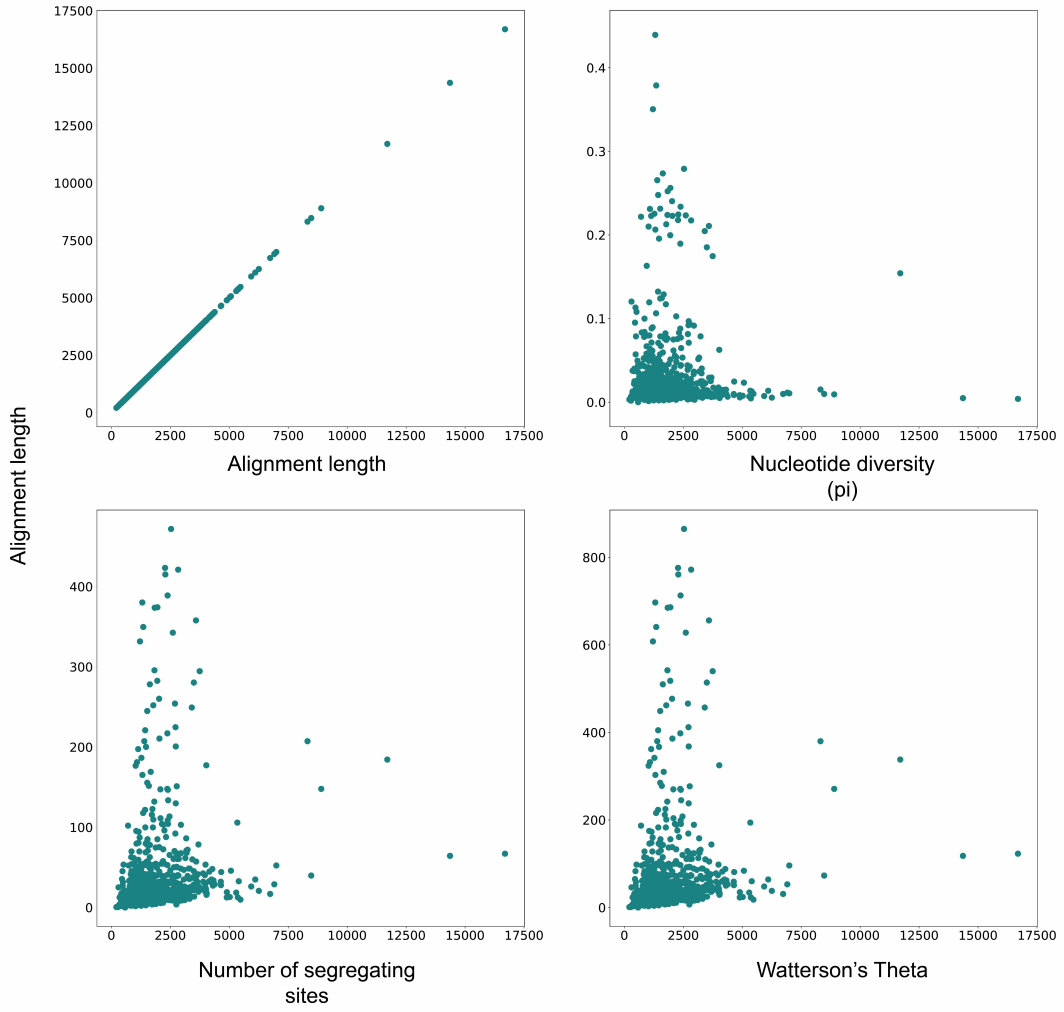


Figure 4.10: Nucleotide diversity, numbers of segregating sites and Watterson's Θ plotted against alignment length for the 759 alignments of the single-copy **BUSCO** genes shared across the four *Tococa* assemblies.

segregating sites (Figure 4.12). Nucleotide diversity, Watterson's Θ and Pi vary according to the number of segregating sites across quartiles, as it is expected as they are estimated based on the number of segregating sites in the alignment (Table D.2 in D). The MCCS tree and the root canal plotted by **DensiTree** (equivalent to a consensus tree) have the same topology across gene categories, except for the MCC tree of super-high diversity genes (Figure 4.13). The two *Tococa*-like specimens from Putumayo are always recovered as sisters, and the *Tococa*-like specimen from Chocó is recovered as sister to the *Tococa* from Antioquia. However, the MCC tree from super-high diversity

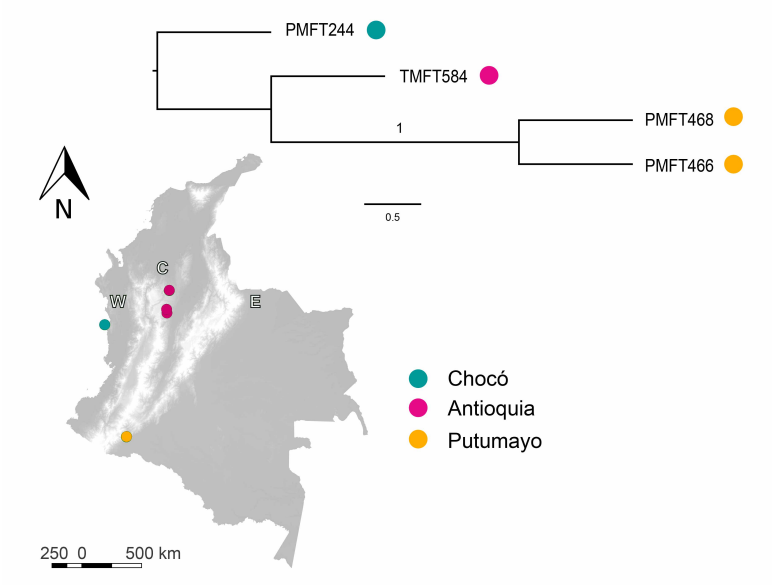


Figure 4.11: Unrooted **ASTRAL** species tree estimated from the 759 alignments of single-copy **BUSCO** genes shared across the four *Tococa* assemblies, color coded by region as shown in the map. The value on the branch corresponds its local posterior probability. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

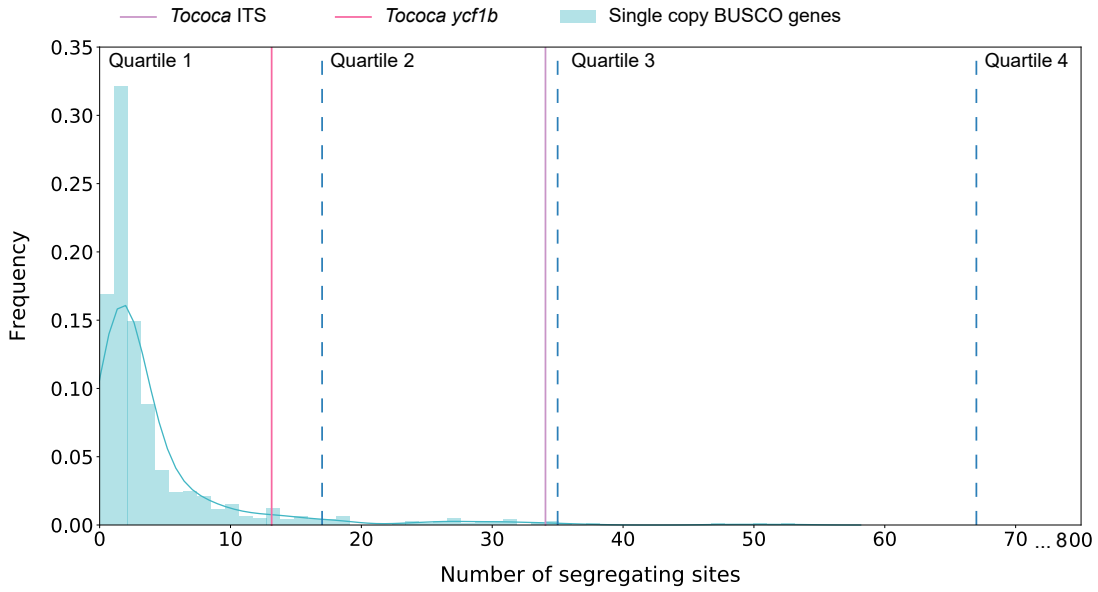


Figure 4.12: Frequency distribution of the number of segregating sites across the 759 single-copy **BUSCO** genes shared across the four *Tococa*-like assemblies and the quartiles of the distribution.

genes recovers Chocó as sister to the Putumayo clade and Antioquia as sister to all of them. Moreover, evidence of gene topology discordance is found in all four categories and is particularly prevalent for the super-high genes. Figure 4.13 shows the most common topology across super-high diversity genes (quartile four, topologies in blue) as an artefact due to **Densitree**'s optimization of tree visualization, but it is still clear that the number of trees with alternative topologies is higher than in other quartiles. For example, the low diversity genes quartile has one tree with the third most common topology, whilst the super-high diversity quartile has four. From all quartiles, medium diversity genes show the highest support for the species tree (the 50 genes are listed in Table D.3 in Chapter D). Additionally, posterior probabilities strongly support the same topology when medium and high diversity genes are used, however, support for the two Putumayo specimens (PMFT466 and PMFT468) is low when low diversity genes are used. When super-high diversity genes are used, high posterior probabilities support the MCCS tree topology, which differs from the consensus root canal and the other quartile topologies.

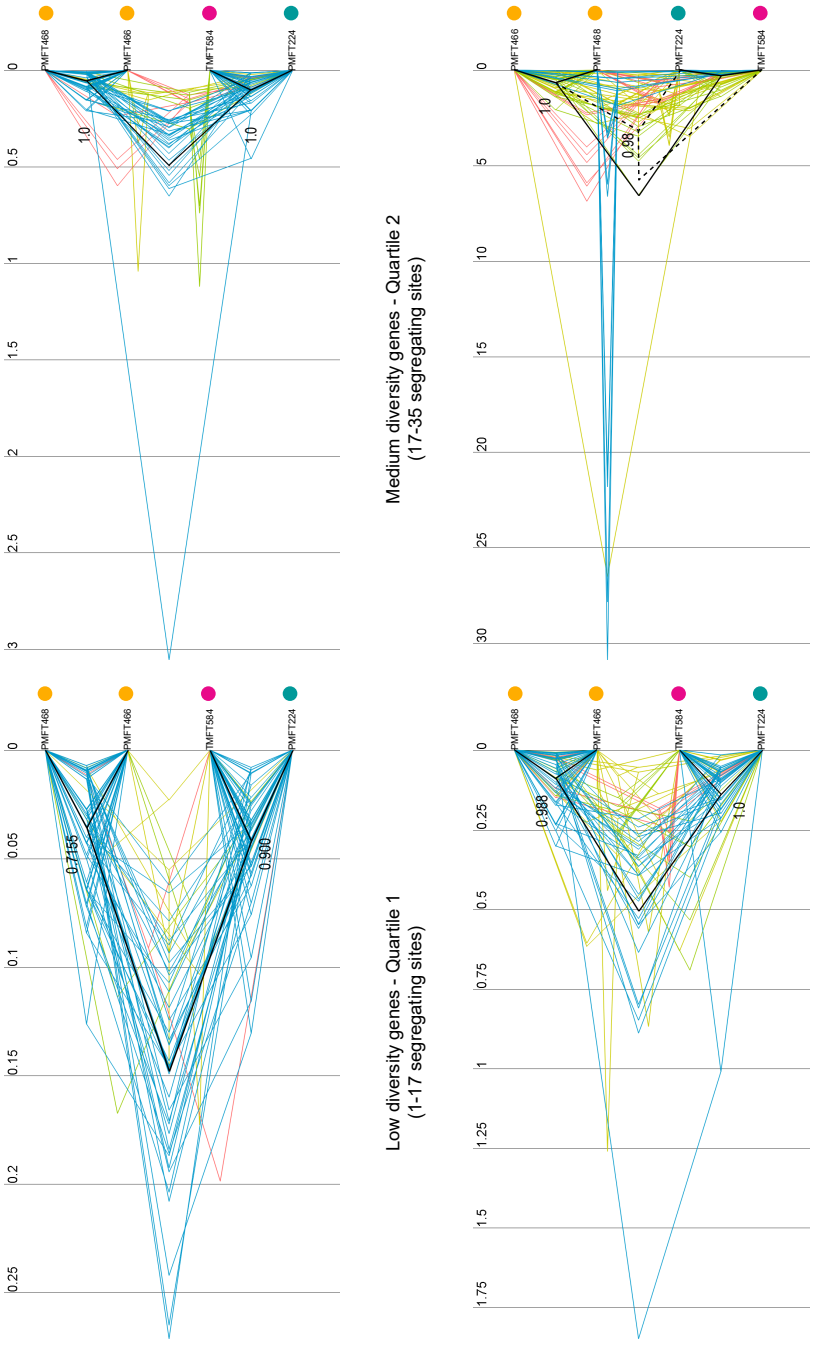


Figure 4.13: MCCS tree and 50 gene trees from the four quartiles of genes based on the number of segregating sites in the alignments. The MCCS tree and the root canal (usually the same) are represented in black. The alternative MCCS tree of the super-high diversity (quartile 4) genes is represented by dashed lines as it differs from **Densitree**'s root canal. Blue represents the first most common topology among gene trees, green represents the second most common, red the third most common and yellow the fourth most common. *Tococa*-like specimens are color coded by region as shown on the map. W= Western Cordillera, C= Central Cordillera, E= Eastern Cordillera.

4.4 Discussion

In this section, I will first discuss the performance of traditional markers in the reconstruction of the relationships between *T. guianensis* and other *Tococa* and Miconieae species and in the identification of the specimens collected across the Andes. Then, I examine the geographic structure of the *Tococa*-like specimens and its relationship with the Andean uplift. Finally, I discuss the assembling process of four *de novo* genomes from four *Tococa* specimens and how these assemblies can be useful to propose new informative loci to improve species tree estimations.

4.4.1 Phylogenetic relationships in *Tococa* and its close relatives

For the *Tococa* specimens collected in this project, the nuclear ITS locus provided more resolution than the two chloroplast markers, especially at the population level. ITS has more segregating sites than the two chloroplast regions, which evolve slowly and provide little information to resolve the relationships between specimens. Studies in other plants show similar results and demonstrated that ITS provides greater resolution than plastid markers especially at low taxonomic levels and when plant species share the same chloroplast haplotype (Acosta and Premoli, 2010; Li et al., 2011). ITS has been proven useful in many instances, such as the identification of alder, palms, and to identify populations of *Notophagus* plants across the Central and Southern Andes (Acosta and Premoli, 2010; Ren et al., 2010; Hollingsworth et al., 2011; Jeanson et al., 2011).

Previous findings on tree topology and the relative position of the genera within Miconieae were confirmed with the ITS and ycf1b trees. First, the main clades within the

Miconieae tribe as described in Goldenberg et al. (2013) are recovered in the ITS and ycf1b trees, but the relationships between them are still poorly resolved. Moreover, an exhaustive phylogenetic study on *Conostegia* including six loci and discrete and continuous morphological characters resulted in patterns very similar to those I found in *Tococa* (Kriebel et al., 2015). It is possible that the lack of information provided by ycf1b and ITS (at the species level) resulted from the rapid radiation of Miconieae and the short time to sort gene lineages across taxa. In a study about diversification rates in Myrtales, Berger et al. (2016) detected a single shift in diversification rates within Miconieae, particularly in the branches leading to *Miconia*, *Clidemia*, and *Tococa*. The increase in diversification rates and the poor resolution within Miconieae regardless of the species sampling are consistent with patterns of a rapid radiation. Second, *Miconia*, *Conostegia*, *Clidemia* and *Tococa* are paraphyletic (Figures 4.1 and 4.2), as previous studies have indicated (Michelangeli, 2000; Michelangeli et al., 2004; Goldenberg et al., 2013). This likely reflects shared polymorphisms that were not sorted during the diversification processes, in addition to phenotypical plasticity.

Morphological identification of the plants was not easy, especially when the individuals were sterile. All the plants used in this project fit within the description of *T. guianensis* in the Flora Neotropica (Michelangeli, 2005, 2010a). The plasticity of the diagnostic traits across the species distribution range and the overlapping of traits across different species hinders the identification process. As collections focused only on *T. guianensis*, specimens collected for this project were expected to form a single monophyletic clade that would not include any sequences from other species. However, this was not the case for any of the phylogenies, and some Amazonian samples fall within the **A** clade (Figure 4.1), which corresponds to the *Clidemia* grade, section *Miconia* according to

(Goldenberg et al., 2013). This suggests that those samples, despite superficially appearing like *T. guianensis* might be another species. Similarly, two reference samples from *T. guianensis* fell into different subclades within the **E** clade (Figure 4.1). It is difficult to know if both subclades correspond to two different lineages of *T. guianensis* or to different species, mostly because of sequences from *T. quadrilateral* and *T. gonoptera* fall in the same subclade with *T. discolor* (synonym of *T. guianensis*) and *Maieta*. Furthermore, most sequences belong to specimens identified as *T. guianensis*, at least morphologically. This proves that the status of species, and even that of the genus, do not match with molecular units for most of the Miconieae tribe. Recently, Michelangeli et al. (2016) have proposed synonymizing *Maieta* and *Tococa* with *Miconia* based on the lack of a stable solution to resolve the tribe's taxonomy using molecular and morphological traits. Complex taxonomy and low phylogenetic resolution seem to be the norm rather than the exception among Miconieae genera. Because *T. guianensis* is not a monophyletic unit, the following sections of the discussion (and this study in general) will treat the specimens as *Tococa*-like specimens.

The correlation between geographic and genetic distances is mainly driven by the *Tococa* from Amazonas in clade **A**. Those sequences belong to a different clade from the rest of *Tococa*-like specimens and are the most distant geographically. Other Amazonian *Tococa* sequences are like sequences from other regions, despite the distance between locations. The lack of genetic differences, mostly ycf1b sequences, reflects the lack of informative sites and the slow evolutionary rates of the chloroplast marker. Nevertheless, the slope of the correlation is shallow and the predicting power of geographic distance for genetic distance is also low. Genetic distance matrices (Figure 4.6) do not show significant differences between populations in the same side or across the Andes with the only exception of the Amazonian *Tococa* from clade **A**. The distance in km between

collecting sites in Amazonas and Meta is approximately 600 km, compared to a distance of 300 km between Meta and Antioquia. Nonetheless, Amazonian specimens in clade **E** are closer to those from Meta than they are to Amazonian specimens in clade **A** and specimens from populations west to the Eastern Cordillera. Based on these differences, a scenario of IBD is less supported than a model of limited gene flow due to the Andes. The role of the Andes is discussed further later in this chapter. Finally, ITS shows that populations in the same side tend to be genetically closer amongst them than to populations across the Andes; however, that tendency is based on few differences.

4.4.2 Time-calibrated phylogeny and phylogeography

Despite *Tococa*'s taxonomic complexity it is possible to make inferences based on the ITS and ycf1b phylogenies and field observations and to propose hypotheses about the geographic structure and the nature of the host-ant interaction. ITS, and ycf1b to some extent, shows an east-west geographic structure consistent with the location of the Andes Cordillera and suggest that the Andes represent a barrier to gene flow between populations on either side. Within the **E** clade in the ITS tree, a western group of *Tococa* derives from an eastern group suggesting that a population already present in the Amazonas, Putumayo and Meta area was isolated by the uplift (Figure 4.14). The fact that the western specimens cluster within the eastern *Tococa* lineage is consistent with an eastern origin with subsequent migration or expansion westward during the early stages of the uplift and before the Andes reached their highest peak (before 20 Mya, Figure 4.4). Similar patterns are found in Neotropical orchids, where lineages of Amazonian origin diversified into Andean and Western lineages because of the Andean uplift and emergence of new niches (Pérez-Escobar et al., 2017). The

Eastern Cordillera in the Northern Andes was already higher than 3000 m.a.s.l. about 10 Ma, and by then it had gone through a period of uplift that accelerated since 23 Mya. The distribution of myrmecophytic *Tococa* species is restricted to below 1200 m.a.s.l. and it is likely that *Tococa*'s habitat was interrupted when the Andes reached a higher height. It is interesting that geographic structure occurs despite the migration events. *T. guianensis* is likely bird dispersed (according to observations in *Miconia* Santos et al. 2017) and one could expect that birds would maintain constant gene flow between populations. But it is possible that by having narrow altitudinal ranges, birds would not always disperse the seeds across the mountains, limiting cross Andean plant dispersal (DuBay and Witt, 2014; Londoño et al., 2017).

The Andes does not always represent a barrier to gene flow and it has proved to be porous in wind-dispersed plants like some orchid species (Pérez-Escobar et al., 2017). However, these mountains represent a continuous barrier to other bird- and mammal-dispersed plants like *Dussia* (Winterton et al., 2014), *Theobroma* (Richardson et al., 2015) and various genera of Annonaceae (Pirie et al., 2006). Despite the phylogenetic split between eastern and western lineages, gene flow between Antioquia (west) and Meta (east) likely persisted during the uplift through areas of low elevation and helped by the presence of continuous patches of forests. Those conditions could have favored the establishment of the plant in the forested areas and the movement of *Tococa* dispersers. However, the time at which it occurs is not clear as the node supports of Meta and Antioquia tips (**c** in Figure 4.5) are not high. In the future, markers adequately selected to perform at population levels can help in testing models of gene flow between these populations.

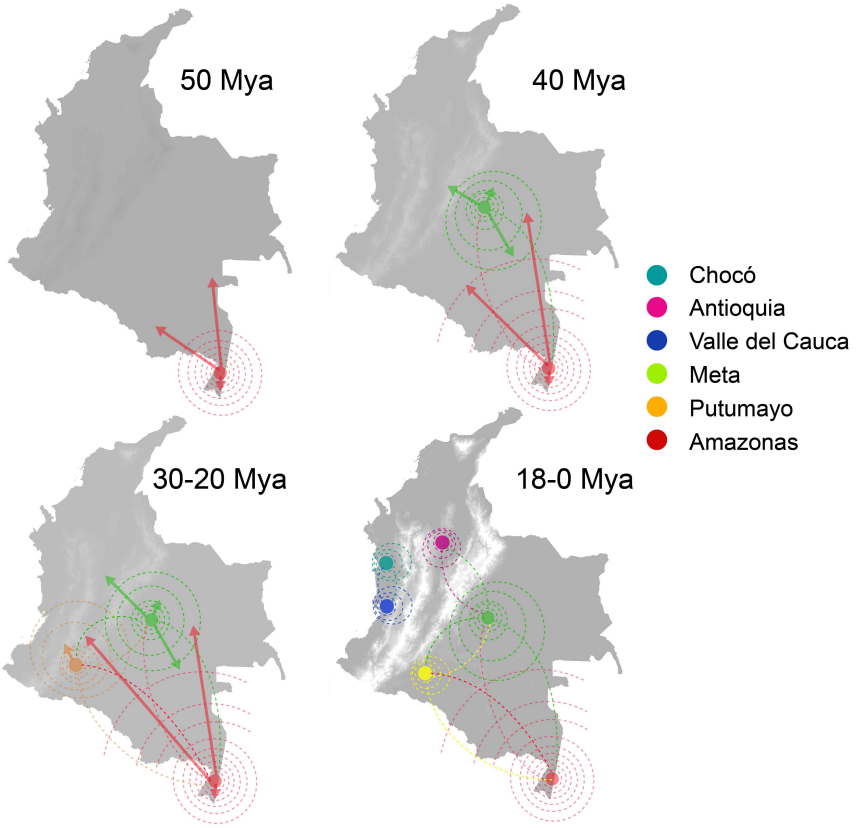


Figure 4.14: Hypothetical scenario for the diversification of *Tococa*-like lineages. Miconieae was already established in South America around 50 Mya, probably in what is now the Amazon basin. That population expanded and diversifying first towards the north-west (between 40-30 Mya) and second, towards the west (between 30-20 Mya). Between 18-10 Mya, lineages continued expanding towards the west as the Andean uplift isolated them and promoted divergence. Gene flow is likely to have continued between Meta and Antioquia populations, at opposite side of the Eastern Cordillera, through an area of relatively low height in the Andes. Collecting locations of the samples are represented by filled circles and the colors represent the geographic location and the transition to new locations. Dashed circles represent areas where gene flow could have continued, and arrows represent directions of population expansion or bidirectional migration.

4.4.3 *de novo* genome assembly of *T. guianensis*

Assembling plant genomes is challenging because of the high occurrence of gene duplications, presence of extensive repetitive sequence regions, low initial DNA quality and quantity, and often high genome size (Bleidorn, 2017). Generation of the *de novo* genome assemblies for *Tococa* specimens was no exception. Perhaps the factors that

most made *Tococa* assemblies challenging were the low initial DNA quantity and quality, duplications, and the many repetitive elements. During the development of this project, whole genome sequencing and library making kits have improved substantially, with technologies that allow for longer reads whilst requiring lower genomic DNA concentrations than before, such as the Nanopore low input kit or the PacBio technologies. But for some organisms, these improvements are not enough. The high content of secondary metabolites in *Tococa*'s leaves (Renner et al., 2001), in addition to damage during collection, contamination and the humidity of the lowland forest fosters DNA degradation. These conditions can increase DNA degradation and thus narrow down the options of read length sequencing. In these cases, the input DNA quality is low and already fragmented into sizes smaller than those needed for long-read sequencing technologies.

The size of *Tococa*'s genome (339 Mbp), although bigger than *Arabidopsis thaliana* (135 Mbp), is much smaller than *Capsicum annum* (3.48 Gbp) and other model plants, and is not on the unmanageable side of the size spectrum. Contrary to expectations derived by the kmer size frequency plots (Figure D.2 in Appendix D), *T. guianensis* assemblies contain a low percentage of complete and duplicated genes. **BUSCO** reports up to 81% completeness of the *T. guianensis* **MetaSPAdes** low-coverage assemblies, from which 4% to 9.65% are complete and duplicated (Figure 4.8). Deeper coverage assemblies of *Arabidopsis thaliana* and *Quercus lobata* analyzed with **BUSCO** result in 93.3% and 90.2% complete respectively. From those, the variation in percentage of complete duplicated genes goes from 37% to 87.2% in *Arabidopsis* and from 37% to 52% *Quercus* (Sork et al., 2016; Jarvis et al., 2017). However, it is important to keep in mind that low coverage assemblies can be missing a large portion of the genomes and the results presented here can only be confirmed with deeper coverage sequencing.

Moreover, **BUSCO** does not report less conservative genes and repetitions suggested by the kmer plots might correspond to non-coding repetitive regions that will nevertheless complicate the genome assemblage.

Once the libraries are sequenced, the challenge is to find a genome assembler capable of dealing with the characteristics of the genome without increasing the memory required to perform the task. Despite choosing assemblers based on their capacity to deal with heterozygosity, repeats, and low coverage, some assemblers did not perform as expected. **Masurca** (Zimin et al., 2013) is designed to deal especially with Illumina data, and not much focus was given to the features of the genomes. **Platanus** (Kajitani et al., 2014) was designed to deal with highly heterozygous diploid genomes but failed to produce large contigs, which reduces the quality of the assembly. **Velvet** (Zerbino and Birney, 2008) is perhaps the most robust assembler, but requires large amounts of memory that, when not available, terminate the assembling process with errors. **MetaSPAdes** (Nurk et al., 2017) produced the longest, most complete assemblies for *Tococa*. Released only in 2017, it implements new algorithms for the effective resolution of repetitive elements, fast construction of assemblies and fast read correction that reduces errors in the reads. Although both sets of libraries were produced to increase the coverage of the genomes, efforts to assemble the genomes using the reads from both library preparations failed (data not shown), possibly because of excessive memory requirements needed to solve ambiguities. To help genome assemblies, it is important not only to be aware of the possible limitations imposed by the genome itself but also to test different tools and protocols that can deal better with the data. Improvements in bioinformatics tools allow the assemblage of genomes of non-model organisms in a faster and more accurate way and will continue to do so as the field progresses.

The *Tococa de novo* assemblies presented here are the first whole genome assemblies for the Melastomataceae family, that represents 160 genera and 4,079 recognized species (The Plant List, `http://www.theplantlist.org/1.1/browse/A/Melastomataceae/`, consulted the 20-09-2017, Chen and Renner 2007). Two transcriptomes and seven mitochondrial genomes had been sequenced, but these miss the information that a whole genome can provide: non-coding regions. The assemblies generated in this project will be used in the future to perform detailed population genomic studies and to assist with the assembly of more *Tococa* specimens and of other species in Melastomataceae. More general applications for whole genome assemblies include Genome-Wide Association Studies (GWAS) and applications of Approximate Bayesian Computation (ABC) approaches to estimate population parameters. GWAS test associations between allele variants and traits (*e.g.* drought resistance, flower coloration) by comparing genomes across populations expressing the polymorphic trait. Applications of ABC approaches include testing models of diversification with gene flow between three populations of the gall wasp *Biorhiza pallida* (Robinson et al., 2014), and speciation in the white lizard genera *Sceloporus* and *Aspidoscelis* (Laurent et al., 2016).

Another application for whole genome data is the search for markers that will help resolve phylogenies of complex taxa. If enough assemblies are available, orthologous regions can be aligned and metrics of those alignments used to guide the selection of candidate regions. Such regions can be selected to have the variation and information appropriate for the type of evolutionary relationships to resolve, *e.g.* among higher taxa, species or populations. Here, I aligned 759 single-copy **BUSCO** genes from four *Tococa* assemblies representing three populations within a single species and calculated the number of segregating sites as a proxy for the phylogenetic variability of the markers. I then reconstructed the species trees using four categories of genes with different numbers

of segregating sites and calculated the posterior probability of the nodes. Those species trees can be compared to the **ASTRAL** tree as it was estimated from independent alignments of the 759 single-copy **BUSCO** genes shared across four *Tococa* assemblies, accounting for the coalescent processes of multiple gene histories (explained in depth in Chapter 3).

The topologies of the ITS tree, *ASTRAL* species trees and three of the Bayesian trees estimated with low, medium and high diversity genes have the same topology. Only the tree estimated using super-high diversity genes had a conflicting topology, likely produced by a high proportion of gene tree discordances. Assuming the same mutation rates, genes with a high number of nucleotide substitutions have larger effective populations sizes, and therefore more ancestral polymorphisms are expected to remain in the population, increasing incomplete lineage sorting between lineages. Medium diversity genes, with 17-35 segregating sites, have the highest node support for the *Tococa* tree. Using super-high diversity genes, the resulting topology differs from the one produced with lower diversity genes possibly due to strong incomplete lineage sorting and an excess of segregating sites that can saturate the alignment. ITS falls within the category of medium diversity genes (quartile two), and alone provided better resolution with a better sampling than ycf1b (which falls within the low diversity gene category, Figure 4.12). This agrees with better resolution provided by genes in the second quartile than other quartiles. The aim of this approach is not to select markers that will have a universal application but to demonstrate how whole genome data can be useful in selecting markers before committing to more sophisticated protocols. In the future, this analysis can be repeated controlling by alignment length and thus, using evolutionary rates instead of only nucleotide diversity. Moreover, this method shows how it is possible to select markers to resolve evolutionary histories within a time frame and

sampling window in mind. Another example of selection of genes to increase resolution in phylogenies is Nicholls et al. (2015). In their paper, they filter and select markers that will likely resolve the relationships between species of the neotropical tree genus *Inga*, which has around 391 recently diverged species (Richardson et al., 2001). After selecting the markers, Nicholls et al. (2015) used target sequencing for those markers and successfully assessed them by resolving between-species relationships. Finally, for my approach to be successful, one must use genomes that maximize the genetic diversity expected among the taxa of interest. Studies can also be improved by testing different outgroups and selecting the most appropriate ones.

Despite the status of *Tococa* as a genus is uncertain, this chapter reveals how *Tococa*-like specimens are geographically structured, provides estimates of the divergence times between *Tococa* populations and generates genomic information valuable for studies on Melastomataceae or on mutualisms. Furthermore, it provides evidence of how evolutionary history of rapid radiations can be difficult to resolve using few loci but also provides tools to select appropriate markers for a targeted taxonomic level in the absence of a close reference genomes and when more data for the organism is not available. Future directions for the genomic data include the use of both libraries to increase genome coverage and generate an appropriate reference annotated genome. Deeper population analyses are possible after the genomes are appropriately aligned, and allele variants are filtered, a process that requires computational time but that confers confidence on the results. Population divergence models that can be tested include the estimation of gene flow events between Meta and Antioquia populations and the origin (dispersal versus panmictic) of the ancestral populations of *Tococa*.

CHAPTER

5

DISCUSSION

5.1 Introduction

Mutualisms are shaped by selective pressures that stabilize the costs and benefits of the association, within a geographic and temporal context. Ant-plant mutualisms (here defined as those in which the ants live inside hollow structures provided by myrmecophyte plants) are restricted to the tropics. These mutualisms are also especially diverse in the Neotropics, an area of recent and rapid geographical change. The prevalence of such diversity in the Neotropics has been attributed to the consequences of the Andean uplift and the change in landscape and climate patterns that resulted in the speciation of plants and animals. Neotropical ant-plant mutualisms are not exempt from the

effect caused by those changes and exploring their responses to them is essential to understanding their evolution.

5.1.1 Identification of the partners

This project explores the role that past changes in Andean orogeny can have in the evolution of mutualisms by looking at the evolutionary patterns of the *Tococa guianensis* plants and their associated *Azteca* ants. To do so, it is necessary to identify which partners are associated across the mutualism's distribution range, the relationships between both partners' lineages and the temporal and geographic context in which the mutualisms developed. The *T. guianensis-Azteca* system is particularly interesting because of its widespread distribution across the Andes. The availability of fossil records allows us to examine the evolution of the system in a temporal context. However, both groups of organisms represent challenges that often deter researchers from studying them. *T. guianensis* is well known for its paraphyly as a species (and generally *Tococa* as a genus). In addition, the identification of both *Tococa* and *Azteca* is challenging as the intraspecific variation of characters, frequently used as diagnostic, is substantial and often overlaps with characters diagnosing of other species.

Chapter 2 identifies the *Azteca* ants inhabiting *T. guianensis* and Chapter 4 evaluates the membership of sampled specimens to the established Linnean species *T. guianensis*. Both chapters use DNA barcodes to identify the specimens, evaluate the geographic structure and produce calibrated phylogenies. For *Azteca*, two major lineages were identified and the conflicting position of a Santander population revealed between ITS2 and COI (Figures 2.6 and 2.7 in Chapter 2). These Eastern and Western *Azteca* show a strong geographic structure and enough genetic divergence that suggest they can be

different species. The divergence time between both coincides with the Andean uplift stage in which the height of the mountains could have already represented a genetic barrier. Less sampled *Azteca* lineages were identified and correspond to singletons or groups of a few sequences from individuals whose membership changes depending on the threshold or method used to delimit the MOTUs. A more complete sampling is needed to correctly assess the membership of those groups but getting enough sampling would require a larger survey of the host plants and the construction of rarefaction curves. This can be difficult because the distribution of the plants inhabited by *Azteca* is poorly known (except for this study). Here, because the identity of the ant colony is unknown before opening the domatia, the sampling protocol targeted *T. guianensis* and not *Azteca*. Besides, the proportion of specimens per *Azteca* lineage will reflect the frequency at which those lineages are naturally found inside *Tococa*.

On the other hand, DNA markers for plant identification was not enough to fully resolve the relationships within Miconieae, but it was enough to place most collected specimens within *T. guianensis* and to observe certain degree of geographic structure (Figures 4.2 and 4.3 in Chapter 4). The lack of resolution in *Tococa* and other Miconieae genera lead to the conclusion that different markers from those traditionally used in plant phylogenetics are needed. Moreover, that finding markers with the level of variability necessary to resolve relationships requires either large amounts of genomic data or significant funding to account for sequencing failures. Hence, the genome assemblies obtained here are of great utility. Not only do genome assemblies allow searching for regions with appropriate levels of sequence variation to resolve a range of phylogenetic problems, but they also contain the conserved flanking regions required to design primers or baits for targeted sequencing experiments. Furthermore, most enrichment and transcriptome sequencing require at least a draft of a genome for the selection

of loci to sequence. An example of the utility of the genome assemblies is presented in Nicholls et al. (2015). There, they use three *de novo* transcriptomes to select and sequence only targeted markers from a larger number of species and resolve the phylogeny of the rapidly radiated *Inga* plants. The advantages I found in using whole genome sequencing instead of transcriptomes are that DNA extraction is easier than RNA extraction, especially if the tissue material is limited or of low quality. It also provides a larger range of informative regions, as the sequencing includes introns and non-coding regions that can have prints of evolutionary processes. The method proposed here selects candidates from single-copy genes across the samples, reducing the chances of introducing paralogues into the analyses. Reciprocal Blast and mapping those genes back to the assemblies using stringent parameters are additional steps to detect paralogues (*e.g.* if more than one similar but not identical regions in the assembly aligned with the single-copy genes). One consideration, however, is that the more divergent the taxa of interest are, the fewer single-copy genes they share and higher the probability of those genes being absent from the taxa used after selecting the markers.

5.1.2 Geographic structure

T. guianensis and *Azteca* lineages have similar geographic structure and their distributions are divided between west and east relative to the Andes Eastern cordillera. Such pattern is stronger in *Azteca* than it is in *T. guianensis*. At a more local level, only the mitochondrial COI in *Azteca* exhibits finer structure coinciding with the collecting sites (Figures 2.5 and 2.6 in Chapter 2). For *Tococa*, the nuclear ITS phylogeny is better resolved than the chloroplast phylogenies and consequently the geographic structure is more evident (Figures 4.1, 4.2 and 4.3 in Chapter 4).

Although overall patterns are alike, direct comparisons between plant and ant geographic structures should be made with caution, as such structure is conditional on the markers used. However, evidence supports the Andes acting as a barrier to gene flow between Eastern and Western *Azteca*, and to some extent between western and eastern *T. guianensis*. The results also demonstrate that the Eastern Cordillera is a more effective barrier than the Central or the Western Cordillera. Those last two did not have a significant effect on structuring either ant or plant lineages and gene flow persists between populations from both sides of them. This is because those Cordilleras are not continuous, and their altitudes decline towards the north end, meaning that dispersal through lowland forest is possible and thus they are not a barrier to organisms restricted to lowland areas. Besides, the average height of the Western and Central cordilleras lies around 2,000 m.a.s.l and it is possible that this height does not represent a strong barrier, making the Cordilleras porous barrier allowing migration (Winterton et al., 2014; Richardson et al., 2015; Pérez-Escobar et al., 2017). On the other hand, the Eastern Cordillera reaches heights around 2,000 to 5,000 m.a.s.l. and its entire range is closed until it joins to the Merida cordillera in Venezuela. As the upper limits of the distribution of *T. guianensis* is 2,500 m.a.s.l. and that of *Azteca* is 2000 m.a.s.l., the height of the Eastern Cordillera is a limitation for both associates (Longino, 1991a; Michelangeli, 2003). Instances were altitude is a limitation for mutualisms have been explored before. Plowman et al. (2017) studied ant-plant association networks in Papua New Guinea and found that at high altitudes the associations become rarer as the temperatures are colder and the partner availability decreases. At higher altitudes, abiotic stress increases and the mutualisms collapse because of lower population sizes and the reduction of mutualistic benefits (Plowman et al., 2017).

The predictions that dispersal differences between ants and plants will have a strong

influence on the effect that the Andean uplift had upon their evolution and that cross-Andean dispersal can occur in plants but not ants are not strongly supported. Results in Chapters 2 and 4 show that cross-Andean gene flow events can occur among populations of *Tococa* with certain frequency and *Azteca*. Although chloroplast markers do not provide enough resolution to make inferences and geographic structure cannot be inferred from polytomies, the more resolved ITS plant data shows strong geographic structure, except for Antioquia specimens (Figure 4.1 in Chapter 4 and Figure D.3 in Appendix D). In the case of *Azteca*, Meta M3 and M4, because COI groups them with Eastern *Azteca* and ITS2 with Western *Azteca*, it is possible that either sexually-biased migration or mitochondrial capture explain the pattern. A similar scenario is possible for the Santander S2 population, although evidence in Chapter 3 shows that Incomplete Lineage Sorting can be an explanation for the conflict in its position.

The Andes have acted as a barrier to dispersal for lineages in other plant families, including Annonaceae, Rubiaceae and Fabaceae (Pirie et al. 2006; Antonelli et al. 2009; Pennington et al. 2010 respectively), but data from orchids and results from *Tococa* and *Azteca* suggest this is not always the case. Pérez-Escobar et al. (2017) inferred the biogeographical history and diversification of two main Neotropical orchid taxa (Cymbidieae and Pleurothallidinae) and found that cross Andean dispersal events did not decrease during the Andean uplift. This means that for some plants the barrier is indeed porous over long timescales. Seed dispersal in *Tococa* differs from Orchids in that fruits are mostly dispersed by birds (Santos et al., 2017), which are often restricted altitudinally and might not disperse the seeds as widely as the wind-dispersed orchid ones (Julliard et al., 2006; DuBay and Witt, 2014; Londoño et al., 2017). compared to other plant-inhabiting ants, *Azteca* has the capacity to disperse large distances (Yu et al., 2001; Bruna et al., 2011). In an experiment involving an undescribed *Azteca*

species inhabiting *T. bullifera*, Bruna et al. (2011) found that *Azteca* has higher capacity for long dispersal events than *Pheidole minuta* and *Crematogaster laevis* (up to 400 m). *Azteca* and *T. guianensis* populations can be exploiting the areas of the mountain ranges that are low enough to allow gene exchange and dispersal. But dispersal is different from successful establishment, and altitude remains a limitation even if long dispersal is possible. Thus, exceptional cross-Andean dispersal of *Tococa* and *Azteca* must have been facilitated by altitude gaps and finally determined by adaptability, establishment capacity and habitat availability.

As mentioned above, two areas of the Eastern Cordillera range are low enough to allow dispersal. The first area is Santander (for further discussion see Chapter 3, arrows near 1 and 2 in Figure 5.1) near the junction between the Colombian and Venezuelan Andes. The connection between both ranges closed approximately 5.2 Mya, and the peaks are on average 2,000 m.a.s.l. or lower (Hoorn et al., 2010), thus the area represents a gap for gene flow between S2, Western and Eastern *Azteca*. The second area where the Eastern Cordillera is low is between Meta and the Magdalena valley, also representing a gap allowing introgression between M3, M4 and Western *Azteca* (3 in Figure 5.1) and between *T. guianensis* from Antioquia and Meta (Figure 4.1 in Chapter 4). Continuous gene flow through those gaps during population divergence can explain the high ILS found between lineages of plants and ants (Chapter 4 and 3). Large population sizes can also explain the ILS, although it was not tested in this study.

It is possible that gene flow continued through those gaps during population divergence until very recently and that it stopped once habitat conditions changed during the last period of Andean uplift. One possibility is that the recent appearance of dry forest in those areas, resulting from the development of rain shadows, had finally closed the

gaps and stopped the gene flow. Collection records and field observations prove that neither *Azteca* nor *T. guianensis* grow in the dry forest making it a final barrier to gene flow. However, estimated ages for the origin of dry forest in this part of the Andes are not known and it is also possible that gene flow continued even after the rise of the dry forest. The genome assemblies generated for both species during my thesis are a promising resource for the identification of large sets of hundreds or thousands of sequence regions suitable for population genomic analysis. Such datasets have the power required to assess support for models of population divergence with or without gene flow, and to model changes in population size (Hearn et al., 2014; Robinson et al., 2014)

5.1.3 The Andean uplift and the establishment of the mutualism

Although east-to-west split events in *Azteca* and *T. guianensis* did not occur simultaneously, they both coincide with one or another period of activity of the Andean uplift during the Miocene. The uplift of the Andes occurred in several steps, but the most significant peaks of height gain occurred 23 Mya during the late Miocene (Hoorn et al., 2010), During the Eastern Cordillera continuing uprising at 11-18 Mya (Hoorn et al., 1995), and until its final closure with the Venezuelan Andes at 5.2 Mya (Hoorn et al., 2010). The mean age of crown *Azteca* is 26 Mya and the mean age of *Tococa*-associated *Azteca* is 19 Mya and the split between Western and Eastern *Azteca* is 15 Mya across models as estimated using ITS and COI in Chapter 2. Similarly, the age of *Tococa*-associated *Azteca* has a mean of 18 Mya based on the genomic data (Chapter 3). For *Tococa*, the crown age of the clade including most *T. guianensis* specimens (**b.** in Figure D.3, Appendix D) is 6-7 Mya and the split of the biggest western lineage (**c.**

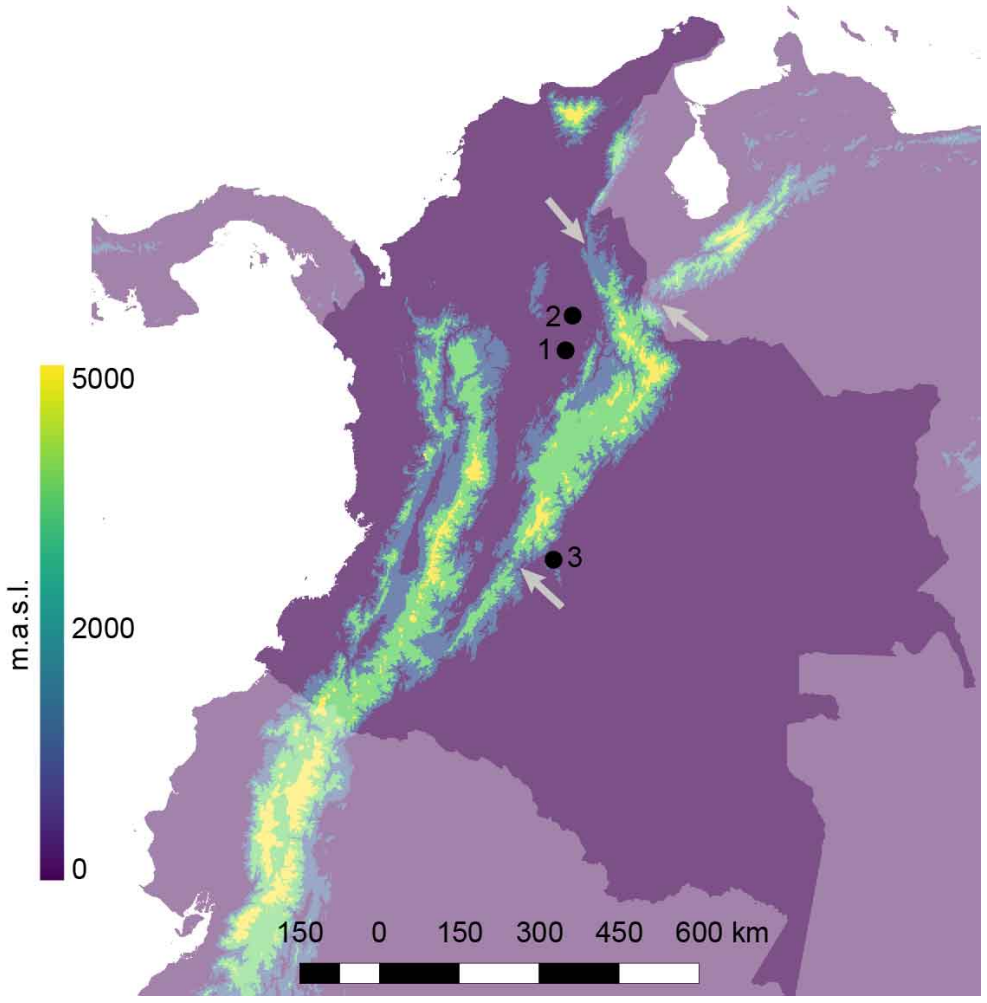


Figure 5.1: Topographic map of the Andes cordillera in Colombia highlighting in greens and yellow areas higher than 2,000 meters above the sea level (m.a.s.l). The grey arrows show areas where cross Andean gene flow between populations of *T. guianensis* and *Azteca* can occur. The location of *Azteca* populations showing conflicting tree topologies are: **1.** Santander-S1; **2.** Santander-S2; **3.** Meta-M3 and M4

in Figure D.3) occurring at around 4.3 My. Likely, dispersal capabilities and flexibility to adapt to higher altitudes could have delayed the barrier effect on *T. guianensis* populations caused by the Eastern Cordillera.

Morley and Dick (2003) proposed a Gondwanan origin for Melastomataceae with the establishment of Miconieae at least 55 Mya in South America and subsequent migration towards North America. However, Berger et al. (2016) propose a much younger origin of stem Miconieae 26-27 Mya and a crown age 20 Mya approximately. As it is briefly

discussed in Appendix D, Berger et al. (2016) estimates are supported by the calibration of the Myrtales order using 10 fossil records. Morley and Dick (2003), on the other hand, rely on calculated substitution rates for their calibrations and therefore their dates are less reliable. Plant diversity in the Amazonian craton peaked in the Miocene around 13 Mya, and despite marine incursions fragmenting the forest 10-20 Mya, the forests around the wetlands already had the same plant family composition it has today (Hoorn et al., 2010), including Melastomataceae species. It is possible that the Neotropical Miconieae tribe originated in the Amazon, and that from there the Amazonian lineage *T. guianensis* crossed from the Amazon, and Putumayo to Chocó and Valle del Cauca, with continuous gene flow through the gaps in the Cordillera (as discussed above). Those migrations (or population expansions) were followed by the isolation of populations, in part associated with the Andes uprising barriers to gene flow.

According to phylogenetic reconstructions, insects diversified and new species emerged in the Amazonian lowlands between approximately 20-12 Mya (Hoorn et al., 2010), some because of the Andean uplift. The diversification of the Neotropical *Melipona* bees (15.4 Mya, Ramirez et al. 2010), and butterflies in the Neotropical clades of the subtribe Euptychiina (between 21 and 15 Mya, (Peña et al., 2010)), are examples of insects other than ants whose diversification is a consequence of the Andean uplift. It is difficult to know where the *Azteca* genus was first established, but one possibility is that it occurred in Central America by 20-15 Mya as evidenced by the fossils from the Dominican amber (Wilson, 1985). *Azteca* likely diverged from *Linepithema* around 66 Mya (Ward et al., 2010). *Linepithema* and *Azteca* are both originally Neotropical (*Linepithema* is best known as the invasive fire ant that has colonized other areas on Earth). *Azteca* and *Linepithema* belong to the Leptomyrmecini crown group, which

diversified in the Neotropics (Ward et al., 2010), but it is difficult to know if the genus was already spread throughout Central and South America or if it migrated and in which direction it migrated. Nevertheless, the data presented here suggest that *Azteca* has been present in Central and South America for longer than *Tococa* has.

5.1.4 Potential host-switches

The *Azteca* ITS2 data shows that *Azteca* lineages associated with *Cecropia* are paraphyletic, with a *Cecropia*-associated *Azteca* clade (**CA**) from Colombia falling within the *Tococa*-associated Western *Azteca* (Figure 5.2). This suggests either one event of host switching from *Tococa* to *Cecropia*, or two shifts from another host to *Tococa* in this *Azteca* clade. Sister species to *Tococa*-associated *Azteca* also associate with other plant hosts. *A. ovaticeps* (AO in Figure 5.2) is a common inhabitant of *Cecropia* and *A. pittieri* and *A. beltii* are usually found in *Cordia*. Thus, it is likely that host switching is a common phenomenon for *Azteca*. Host switching is common in other systems, and such is the case between *Pseudomyrmex* ants and its hosts (Chomicki et al., 2015), *Crematogaster* (Decacrema) ants and *Macaranga* plants (Feldhaar et al., 2003), and between Neotropical figs and fig wasps (Machado et al., 2005). However, the data obtained here does not represent a comprehensive sampling of all *Azteca* species (those inhabiting with plants and those that do not) and thus is not sufficient to reconstruct the ancestral host character or infer the emergence of ant-plant associations in *Azteca*.

Azteca and *T. guianensis* likely had similar ancestral distribution areas at least since *Tococa* species started diversifying (around 16 Mya as shown in Figure D.3 in Appendix D) and at least to the east of the Eastern Cordillera. Chomicki and Renner (2015) suggests that myrmecophytism in the Neotropics dates to 20 Mya. However, it is

unclear if *Azteca* and *T. guianensis* occupied the same niches or even if their association was established or not at that time. Evidence from the *Azteca* ITS2 calibration in Figure 2.13, Chapter 2 suggests that host switches could have also occurred between other plant genera and *Tococa*. After the east-to-west split, *Azteca* could have jumped from a *Tococa* host to a *Cecropia* host after Western *Tococa* diverged. Moreover, the most common ant genus associated with *T. guianensis* in Chocó (west to the Eastern Cordillera) is *Pheidole*, while other *Azteca* species associate with *Cecropia* in that same area. Therefore, host switches are not unlikely at any point of the plant-ant *Azteca* evolution and diversification, testing for this requires a better sampling of non-plant and plant-inhabiting *Azteca* and the reconstruction of the ancestral host associations throughout the complete *Azteca* phylogeny.

Unlike specialist mutualisms where partners cannot migrate separately, *Azteca* could have migrated anywhere if there were other hosts available. According to palynological records, pollen from *Cecropia*, the myrmecophyte *Duroia*, lowland *Miconia* and montane Melastomataceae were present in the Amazonian lowland forest during the Neogene (23-3.6 Mya, Van der Hammen 1956), and during the Late Miocene to Early Pliocene (7-3 Mya). *Miconia* pollen was also predominant there during that period (Jaramillo et al., 2009). Moreover, Chomicki et al. (2015) suggest *Azteca* as a potential generalist associate of *Triplaris* before *Pseudomyrmex* became *Triplaris*' obligate partner. This suggests that multiple potential hosts were present, perhaps sympatric, in the lowland forest before and during the Andean uplift (as well as now).

From all 4079 accepted Melastomataceae species, 84 (within 11 genera) are myrmecophytes, including *Tococa*, *Clidemia*, *Conostegia* and *Henrrietella* among others (Michelangeli, 2010a). In the case of myrmecophyte Miconieae, their status as genera and species

is unresolved. Without a fully resolved phylogenetic hypothesis, it is not possible to infer the number of times myrmecophytism has evolved in Miconieae. However, the sparse distribution of myrmecophyte Miconieae species across the most complete phylogeny published by Goldenberg et al. (2008) suggests that the mutualism could have evolved independently more than once. This is also the case in other ant-plant mutualisms where the ants are obligate symbionts of their hosts, as for *Pseudomyrmex* ants on *Vachelia* (synonym to *Acacia*, Fabaceae). Mutualistic relationships with plants evolved independently twice in *Pseudomyrmex ferrugineus* and *P. nigrocinctus*, both inhabitants of *Vachelia*, proving that mutualisms can evolve repeatedly for the same lineages of ants and plants (Ward and Branstetter, 2017).

5.1.5 The specificity of the *T. guianensis-Azteca* mutualism

Even though ant inhabitants were collected from inside the plants at the same location, ant-to-plant lineage correspondence is not expected nor always observed. The mutualism between *T. guianensis* and *Azteca* ants can be considered as obligate (one depends on the other) but generalist mutualism (there is no species specificity). The interaction between *T. guianensis* and *Azteca* involves feeding of the ants with glandular trichomes and tended coccids, in addition to the plant's absorption of Nitrogen from the colony's waste (Davidson and McKey, 1993; Cabrera and Jaffé, 1994). The *Azteca bequaerti* ants inhabiting *T. guianensis* respond to leaf vibrations as intruders arrive on the host and proceed to expel them, showing territorial aggressiveness behaviors to protect their host (Dejean et al., 2008). In exclusion experiments on the field, Michelangeli (2003) demonstrated that *Azteca* inhabiting ants significantly reduce the effects of herbivory in *T. guianensis* plants. In an ant-plant facultative mutualism

the ants patrol the host and remove potential herbivores. The plant provides nectar usually from extra floral nectaries (EFN) that the ants harvest; but opposite to an obligate mutualism, the plants do not provide housing to the ant colony (Davidson and McKey, 1993; Bixenmann et al., 2011). For instance, ants visit *Inga* plants to obtain the nectar the plant produces in the younger leaves but the colony does not inhabit the plant. The ants have aggressive behaviors against potential herbivores and protect the vulnerable leaves from which they harvest the nectar, nevertheless, the ants do not induce the production of nectar by the plant (Bixenmann et al., 2011). Because *Tococa*-associated *Azteca* colonies inhabit *Tococa* domatia, the mutualism is considered obligate rather than facultative. Furthermore, *Azteca* colonies inhabiting plant hosts were not observed to live or patrol the ground but only the plant. Food dependency, inhabitance restricted to plant structures, and protective behaviors are indicators of an obligate ant-plant mutualism.

In the case of *Tococa* and *Azteca*, species specificity is higher at the lineage level than at the species level. At the species and genus level, the mutualism is generalist because *T. guianensis* and *Azteca* can both associate with other ant and plant genera. *Tococa guianensis* hosts *Pheidole, Dolichoderus, Camponotus* and *Allomerus* ants in addition to *Azteca* (Alvarez et al., 2001; Bizerril and Vieira, 2002; Bartimachi et al., 2015). Similarly, *Azteca* species can associate with *Cecropia* and *Cordia* (Davidson and McKey, 1993; Pringle et al., 2012).

At the lineage level, neither *Tococa* nor *Azteca* show high species specificity to each other and *Azteca* lineages appear to be more specific than *Tococa* lineages. Network and species specificity analyses (Figure 5.3 and Appendix E) show that the *Azteca* lineages collected from *T. guianensis* are also present in other *Tococa* species that are

nevertheless within that same clade; however, these *Tococa*-associated *Azteca* lineages are not associated with other plant genera. On the other hand, *T. guianensis* and the *Tococa* species within that clade associate with the two Eastern and Western *Azteca* lineages, and at least one *Pheidole* species. The *T. guianensis* clade also associates with other ant genera, including opportunistic ants like *Solenopsis* and *Tapinoma*. Another exception to the lineage specificity is a specimen of *T. guianensis* collected in the west and grouping with the east1 *T. guianensis* clade that associates with Western *Azteca*.

Ant and plant lineages can be coevolving tightly with the help of geographic structure and population isolation in ants, but the coevolutionary mechanism is more likely to be diffuse than pairwise (see 1.1 for the definition). In diffuse coevolution the mutualism is subjected to a variety of evolutionary pressures that are not coming from a single lineage of partners but from multiple ones. Diffuse coevolution is also a pattern consistent with host switches, as low species specificity can facilitate jumps to other hosts. Even though the focus of this study is not the estimation of host switch events, the low species specificity and the associations between *Azteca* and several host species suggest that host switches could have taken place throughout *Azteca*'s evolution.

Finally, the results from this study are not conclusive regarding coevolution and specificity between *Tococa* and *Azteca*, but they offer an initial understanding of the system. The sampling for this study targeted *T. guianensis* specimens and that means that the sampling of the plant lineages is better than the sampling of the ants. Because this does not allow to test whether Eastern and Western *Azteca* inhabit plants other than *T. guianensis* or other Miconieae myrmecophytes, the estimation of ant species specificity can be biased. More detailed surveys of all the possible associating lineages within *Azteca* and *Tococa* and experiments on the fitness of those specific associations

are required to draw better conclusions.

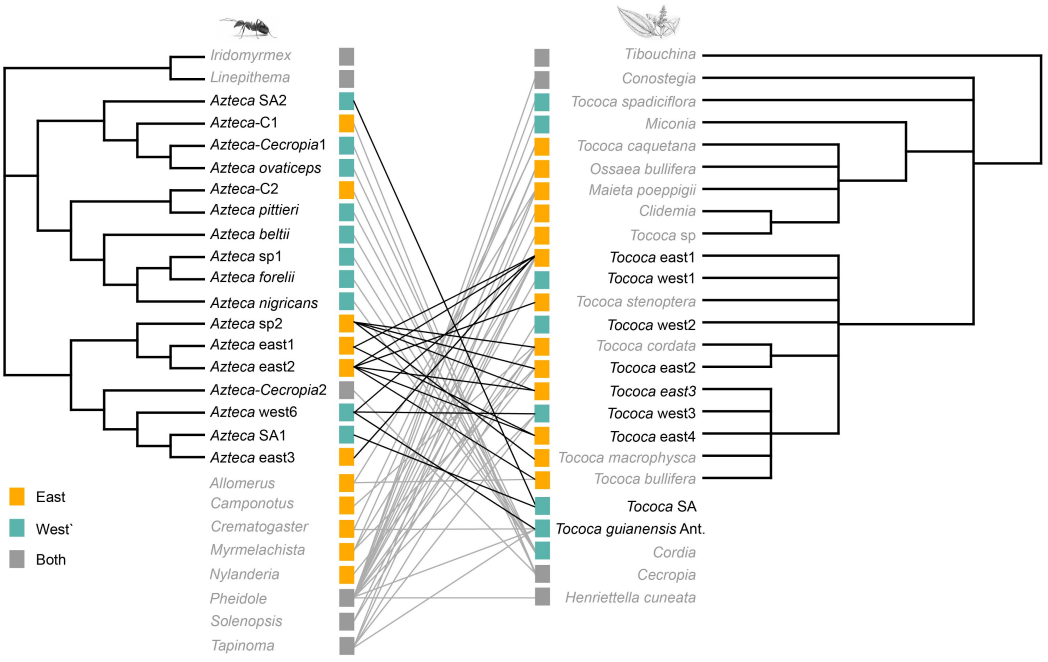


Figure 5.2: Tanglegram showing *Tococa* (left) and *Azteca* (right) ITS calibrated phylogenies and the associations among accessions in the two taxa. Tips are collapsed to the population level and colors indicate the geographic origin of the specimens. Dotted tips represent *Cecropia*-associated *Azteca* from Colombia. AO= *A. ovaticeps*, AP= *A. pittieri*, AB= *A. beltii*. Ant image belongs to Alex Wild.

5.1.6 Ant and plant diversification

The patterns of divergence are different between *Azteca* and *T. guianensis*. In *Azteca*, the Eastern and Western lineages split resulted into two different clades. In *T. guianensis*, the western lineage is nested within an eastern lineage, which is in turn nested within an eastern lineage. There is also more gene flow across the Andes in plants than in ants as is shown by some interleaving between western and eastern *T. guianensis* specimens within the same lineages (Figure 4.5 in Chapter 4 and Figure B.1 in Appendix D). Thus, despite a generalized east-to-west pattern of divergence in both plants and ants, it is unlikely that *T. guianensis* and *Azteca* codiversified in response to reciprocal pressures only.

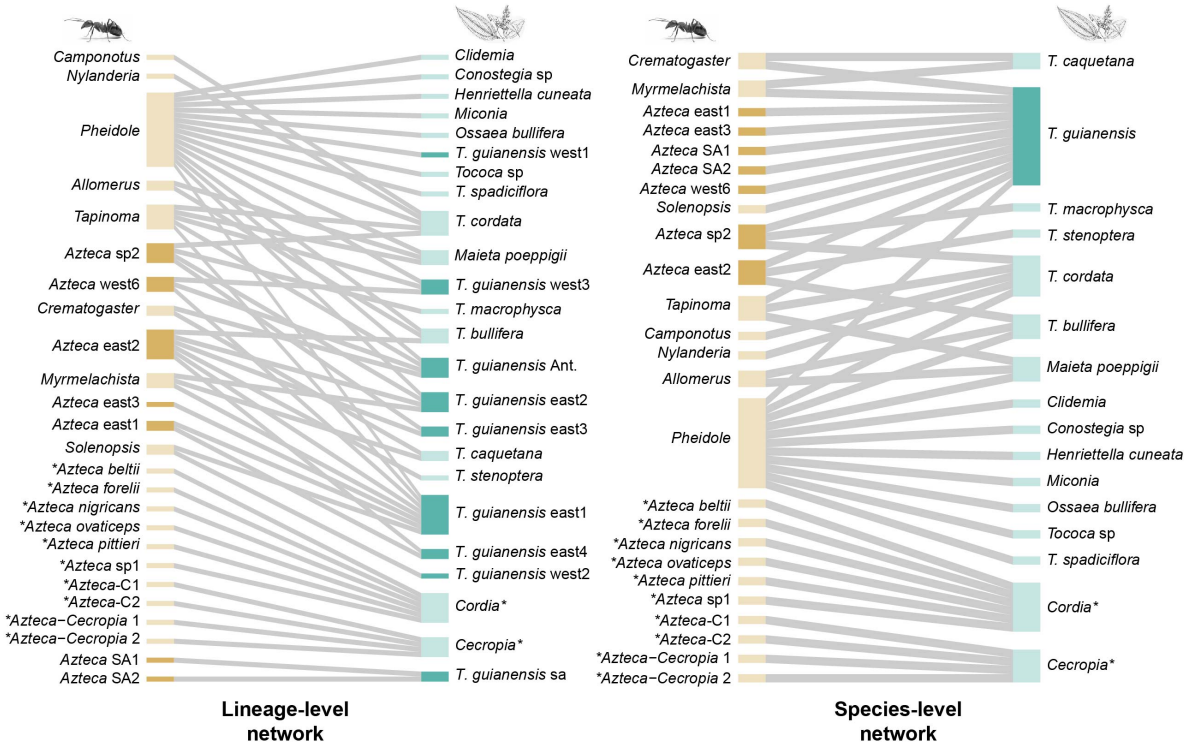


Figure 5.3: Bipartite networks of the ant-plant interactions recorded during this study and the ant-plant interactions reported for the NCBI sequences (*). Bold colors represent the *Azteca* and *T. guianensis* lineages collected during this study and transparent colors represent other interactions. The lineage-level network is drawn from a matrix that differentiates between the different lineages within *T. guianensis*, while the species-level network is drawn from a matrix that clusters all those lineages into *T. guianensis*. Ant image belongs to Alex Wild.

To assess the patterns of codiversification (or the lack of) at the genus level, it is necessary to increase the sampling of both *Azteca* and *Tococa* to include species that are not involved in the mutualism. The collecting sites were chosen to test whether the Andes Cordillera represents a barrier to gene flow between populations or not. Ant-plant associations in the areas that were not sampled are unknown and statements about codiversification are only hypotheses based on the data obtained during this study. Additionally, due to sequencing performance and lack of information provided by the markers, not all the plant samples were sequenced and thus, not all *Tococa-Azteca* associations are depicted. However, those that are shown are sufficient to show the patterns described above.

5.1.7 Ant and plant coevolution

My data provide evidence that ant and plant populations in the *Tococa-Azteca* system became isolated through vicariance and that the shared east-to-west pattern is a consequence of both taxa experiencing population isolation due to the Andes Cordillera. Myrmecophyte *Tococa* and its *Azteca* ants are not the only mutualists to share a history of vicariance. In Borneo, the myrmecophyte *Macaranga* and its *Crematogaster* (Decacrema) ants show the same patterns of population isolation and intraspecific geographic structure probably linked to the rise of the Crocker Mountain Range during the Miocene or to rainforest fragmentation occurring during the Pliocene and Pleistocene (Feldhaar et al., 2003; Banfer et al., 2006). However, codiversification (or cospeciation when occurs at the species level) and covicariance are not mutually exclusive and can explain the observed patterns of evolution on varying degrees. Chomicki et al. (2015) calibrated the phylogenies of *Pseudomyrmex* ants (both plant-inhabiting and non-plant inhabiting species) and the five main host plant groups and found evidence of cospeciation among the youngest nodes, but no evidence of cospeciation and matching divergence times at the genus level. Unlike *Pseudomyrmex*, I found evidence of matching divergence patterns between western and eastern lineages of *Azteca* and *T. guianensis* that are congruent in time with different uplift peaks triggering vicariance, as illustrated by Donoghue and Moore (2003). Besides, hosts switches within Andean *Azteca* and the multiple independent evolution of plant-association within it suggests that codiversification due to coevolution is unlikely at the species level.

Another mechanism that could explain the patterns observed between *Tococa* and *Azteca* lineages is coevolution (*e.g.* reciprocal adaptations in two or more species of organisms resulting from the reciprocal relationship between them, Ehrlich and Raven

1964); however, this is not formally addressed in this project. Coevolution does not require strict codiversification and requires appropriate ancestral trait reconstructions and experiments testing the linkage between those traits and the mutualism. However, coevolution is not rare and there are examples of it occurring between host plants and ant symbionts. One example of coevolution is the *Macaranga-Crematogaster* system. The ant species composition of this system is often determined by the presence or absence of a wax band at the on the surface of *Macaranga* stems at the base of the plant, which excludes ants without the necessary adaptations for retaining grip. Feldhaar et al. (2003) and Fiala et al. (1999) suggests that strict cocladogenesis (or strict codiversification) is unlikely, but that a coevolutionary pattern could exist because ants that cannot run on the wax (typically *Crematogaster* species) inhabit non-waxy *Macaranga* species and those which can run on wax inhabit waxy species. A similar pattern could be important for the *Azteca*-host associations. I observed in the field that *Azteca* species inhabiting *Tococa* are small compared to those inhabiting *Cecropia*, although *Cecropia* sometimes associates with small *Azteca*. A hypothesis is that *Tococa* is imposing a limitation to the size of ants (or even to the colony size); consequently, host shifts of small *Azteca* species from *Tococa* to *Cecropia* would be more likely than large *Azteca* species switching from *Cecropia* to *Tococa*. Hosts species with large caulinary domatia (*e.g.* *Cecropia* species) can host large colonies and large *Azteca* species, while host species with comparatively smaller domatia and prostomas select for associations with small *Azteca* species (*e.g.* *Tococa*, *Duroia* and *Cordia*). If the size of ants, colonies and domatia are traits relevant to the mutualism, coevolutionary mechanisms could give rise to the patterns of evolution between *Tococa* and *Azteca*. However, this pattern can emerge from ecological sorting and detailed ecological surveys are needed to test both hypotheses.

5.2 Conclusions

The Andean uplift had a determinant effect on the evolution of the *T. guianensis-Azteca* mutualism by promoting lineage divergence and limiting gene flow between populations across the mountains. Moreover, the effects of the Andean uplift are greater for *Azteca* ants than they are for *T. guianensis* plants. The congruence between population divergence times and the main peaks of the Andean uplift, in addition to the east-to-west population structure, are expected under scenarios of shared vicariance. The *T. guianensis-Azteca* mutualism is a horizontally transmitted one and coevolution and shared vicariance are important mechanisms that drive its evolution. Nevertheless, reciprocal codiversification due to coevolution cannot be completely ruled out. Hypotheses of coevolution remain to be properly explored, but the availability of genome assemblies makes possible to integrate evolution, genomics and ecology in future studies. Coevolution involves reciprocal changes because of reciprocal relationships (Ehrlich and Raven, 1964). Reciprocal changes along the genomes of associated ants and plants can be explored. After controlling for generation times and other species-specific variables, it is possible to detect regions under selection and, with an appropriate annotation of gene functions, select those that are likely interacting reciprocally. For instance, plant volatile production and ant volatile receptors. Then, it is possible to contrast the evolutionary history of those regions, testing for time and topological congruence. It would be interesting to explore differences in plant volatile production (volatiles are key in ant-plant communications, Jürgens et al. 2006; Dáttilo et al. 2009) and ant defense responses between west and east populations and test if there is selection occurring in this now isolated populations.

Despite the recent population divergence and the prevalence of incomplete lineage sorting in both plants and ants, there is evidence that such divergence coincides with times at which the height of the Andes becomes a limitations to dispersal of *Azteca* and *T. guianensis*. Similarly, there is evidence that both shared a distribution in the past. Whether if the mutualism was established by the time of the population divergence is difficult to know. Based on phylogenetic calibrations of all myrmecophyte plant families and symbiont ants, Chomicki and Renner (2015) suggest that myrmecophytism appeared in the Neotropics around 20 Mya. That date overlaps with the occurrence of the Andean uplift and the divergence of *Azteca* lineages, therefore, is not possible to attribute the diversification of lineages to either the Andean uplift or to the evolution of the associations. Moreover, it is possible that diversification occurred because of both. In the future, the selection of appropriate markers from the genomic data generated in this project will allow for more resolution in the reconstruction of ancestral characters associated with hosts and distribution. Finally, completing the sampling of the Santander populations (and the *Azteca* species in general) will increase the robustness of population analyses and will clarify the demographic history of the lineage.

Genomic data is of great utility to reveal patterns of evolution and to understand the mechanisms behind them. Generating genome assemblies for non-model species is a crucial first step in many applications, including marker screening, population genomic analyses, enrichment sequencing, genome-wide association studies, and (in concert with transcriptome assemblies) understanding the roles of candidate. Furthermore, when lineages are young, and the divergence events are recent, small numbers of loci cannot resolve the evolutionary histories of lineages or their associations. Consequently, the use of genomic data is becoming increasingly important for the understanding of the mechanisms driving the mutualism's evolution.

Future applications of the data presented here include the estimation of population parameters like ancestral population sizes, direction and magnitude of gene flow, and divergence times using allele frequency information from the assemblies. Those parameters, when compared to simulated data under migration and divergence scenarios, allow for testing the fit of alternative models of evolution of ant and plant populations. In addition, targeted enrichment sequencing experiments and results can improve the phylogenetic resolution and better guide the proposal of taxonomic entities; the *Azteca* and *Tococa* genome sequences assembled during this project will serve as a basis for such applications. The same genome assemblies can be used as guides for marker selection under user-selected criteria to resolve the challenging phylogenetic relationships typical of rapid and recent (or even ancestral) radiations.

BIBLIOGRAPHY

Acosta MC, Premoli AC. 2010. Evidence of chloroplast capture in south american *Nothofagus* (subgenus nothofagus, nothofagaceae). *Molecular Phylogenetics and Evolution*. 54:235–242.

Aker C, Udovic D. 1981. Oviposition and pollination behavior of the yucca moth, *Tegeticula maculata* (lepidoptera: Prodoxidae), and its relation to the reproductive biology of *Yucca whipplei* (agavaceae). *Oecologia*. 49:96–101.

Akino T, Terayama M, Wakamura S, Yamaoka R. 2018. Intraspecific variation of cuticular hydrocarbon composition in *Formica japonica* motschoulsky (hymenoptera: Formicidae). *Zoological Science*. 19:1155–1165.

Al-Nakeeb K, Petersen TN, Sicheritz-Pontén T. 2017. Norgal: extraction and de novo assembly of mitochondrial dna from whole-genome sequencing data. *BMC Bioinformatics*. 18:510.

Almeda F. 1997. Chromosome numbers and their evolutionary significance in some neotropical and paleotropical melastomataceae. *BioLlania ed. especial.* pp. 167–190.

Almeda F, Chuang TI. 1992. Chromosome numbers and their systematic significance in some mexican melastomataceae. *Systematic Botany.* 17:583–593.

Alonso LE. 1998. Spatial and temporal variation in the ant occupants of a facultative ant-plant. *Biotropica.* 30:201–213.

Althoff DM, Segraves KA, Johnson MTJ. 2014. *Testing for coevolutionary diversification: linking pattern with process.* 29:82–89.

Althoff DM, Segraves KA, Smith CI, Leebens-Mack J, Pellmyr O. 2012. *Geographic isolation trumps coevolution as a driver of yucca and yucca moth diversification.* 62:898–906.

Alvarez G, Armbrecht I, Ulloa-Chacón EJHAP. 2001. Ant-plant associations in two *Tococa* species from a primary forest in colombian choco (hymenoptera: Formicidae). *Sociobiology.* 38.

Alvarez JM, Hoy MA. 2002. Evaluation of the ribosomal its2 dna sequences in separating closely related populations of the parasitoid *Ageniaspis* (hymenoptera: Encyrtidae). *Annals of the Entomological Society of America.* 95:250–256.

Andrews S, et al. (2 co-authors). 2010. *Fastqc: a quality control tool for high throughput sequence data.* .

Angelis K, Dos Reis M. 2015. The impact of ancestral population size and incomplete lineage sorting on bayesian estimation of species divergence times. *Current Zoology.* 61:874–885.

Antonelli A, Nylander JAA, Persson C, Sanmartín I. 2009. Tracing the impact of the andean uplift on neotropical plant evolution. *PNAS*. 106:9749—9754.

Antonelli A, Quijada-Mascareñas A, Crawford AJ, Bates JM, Velazco PM, Wüster W. 2010. Molecular studies and phylogeography of amazonian tetrapods and their relation to geological and climatic models. *Amazonia, landscape and species evolution: a look into the past*. pp. 386–404.

Antonelli A, Sanmartín I. 2011. Why are there so many plant species in the neotropics?

Araujo SBL, Braga MP, Brooks DR, Agosta SJ, Hoberg EP, von Hartenthal FW, Boeger WA. 2015. Understanding host-switching by ecological fitting. *PLOS ONE*. 10:e0139225.

Arnold ML. 2004. Transfer and origin of adaptations through natural hybridization: were anderson and stebbins right? *The Plant Cell*. 16:562–570.

Ashmead WH. 1905. A skeleton of a new arrangement of the families, subfamilies, tribes and genera of the ants, or the superfamily formicoidea. *The Canadian Entomologist*. 37:381–384.

Aublet J. 1775. Histoire des plantes de la Guiane Françoise. 1. P.F. Didot, London, Paris.

Ayala FJ, Wetterer JK, Longino JT, Hartl DL. 1996. Molecular phylogeny of *Azteca* ants (hymenoptera: Formicidae) and the colonization of *Cecropia* trees. *Molecular Phylogenetics and Evolution*. 5:423–428.

Bacon CD, Silvestro D, Jaramillo C, Smith BT, Chakrabarty P, Antonelli A. 2015. Biological evidence supports an early and complex emergence of the isthmus of panama. *Proceedings of the National Academy of Sciences*. 112:6110–6115.

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet.* 8:619–631.

Baker CCM, Martins DJ, Pelaez JN, Billen JPJ, Pringle A, Frederickson ME, Pierce NE. 2017. Distinctive fungal communities in an obligate african ant-plant mutualism. *Proceedings of the Royal Society of London B: Biological Sciences.* 284.

Ballard JWO, Whitlock MC. 2004. The incomplete natural history of mitochondria. *Molecular Ecology.* 13:729–744.

Banfer G, Moog U, Fiala B, Mohamed M, Weising K, Blattner FR. 2006. A chloroplast genealogy of myrmecophytic *Macaranga* species (euphorbiaceae) in southeast asia reveals hybridization, vicariance and long-distance dispersals. *Molecular Ecology.* 15:4409–4424.

Bao E, Jiang T, Girke T. 2014. Alignngraph: algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics.* 30:i319.

Barriga PA, Dormann CF, Gbur EE, Sagers CL. 2015. Community structure and ecological specialization in plant–ant interactions. *Journal of Tropical Ecology.* 31:325–334.

Bartimachi A, Neves J, Vasconcelos HL. 2015. Geographic variation in the protective effects of ants and trichomes in a neotropical ant–plant. *Plant Ecology.* 216:1083–1090.

Bascompte J, Jordano P. 2007. Plant-animal mutualistic networks: the architecture of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* 38:567–593.

Beattie A. 1989. *Myrmecotrophy: Plants fed by ants.* 4:172–176.

Beattie AJ. 1985. The evolutionary ecology of ant-plant mutualisms. Cambridge University Press.

Beattie AJ, Hughes L. 2002. Ant–plant interactions. *Plant-animal interactions: An evolutionary approach*. pp. 211–235.

Bell CD, Donoghue MJ. 2005. Phylogeny and biogeography of valerianaceae (dipsacales) with special reference to the south american valerians. *Organisms Diversity & Evolution*. 5:147–159.

Beltrán M, Jiggins CD, Bull V, Linares M, Mallet J, McMillan WO, Bermingham E. 2000. Phylogenetic discordance at the species boundary: Comparative gene genealogies among rapidly radiating *Heliconius* butterflies. *Molecular Biology and Evolution*. 19:2176–2190.

Bensasson D, Zhang DX, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in ecology & evolution*. 16:314–321.

Bentham G. 1840. Enumeration of plants collected by mr. schomburgk in british guiana. melastomataceae. *Journal of Botany (Hooker)*. pp. 286–315.

Bentham G. 1844. Melastomataceae. Botany of the voyage of H.M.S. Suplhur. Smith, Elder & Co., London.

Bentham G. 1845. Plantae Hartwegianae. W. Panplin, London.

Berger BA, Kriebel R, Spalink D, Sytsma KJ. 2016. Divergence times, historical biogeography, and shifts in speciation rates of myrtales. *Molecular Phylogenetics and Evolution*. 95:116–136.

Bickford D, Lohman DJ, Sodhi NS, Ng PK, Meier R, Winker K, Ingram KK, Das I. 2007. Cryptic species as a window on diversity and conservation. *Trends in Ecology & Evolution*. 22:148–155.

Bitallion C. 1982. Aspects morphologiques et biologiques de deux espéces de melastomatacas myrmecophiles guyano-arnazonie nnes: Maietaguianensis aublet. *Tococa guianensis Aublet. Theses. Paris. 120pp.* .

Bixenmann RJ, Coley PD, Kursar TA. 2011. Is extrafloral nectar production induced by herbivores or ants in a tropical facultative ant–plant mutualism? *Oecologia*. 165:417–425.

Bizerril MX, Vieira EM. 2002. Azteca ants as antiherbivore agents of *Tococa formicaria* (melastomataceae) in brazilian cerrado. *Studies on Neotropical Fauna and Environment*. 37:145–149.

Blaimer BB. 2012. Untangling complex morphological variation: taxonomic revision of the subgenus *Crematogaster* (oxygyne) in madagascar, with insight into the evolution and biogeography of this enigmatic ant clade (hymenoptera: Formicidae). *Systematic Entomology*. 37:240–260.

Blatrix R, Debaud S, Salas-Lopez A, Born C, Benoit L, McKey DB, Attéké C, Djiéto-Lordon C. 2013. Repeated evolution of fungal cultivar specificity in independently evolved ant-plant-fungus symbioses. *PLOS ONE*. 8:1–9.

Blattner FR, Weising K, Bänfer G, Maschwitz U, Fiala B. 2001. Molecular analysis of phylogenetic relationships among myrmecophytic *Macaranga* species (euphorbiaceae). *Molecular Phylogenetics and Evolution*. 19:331–344.

Blaxter ML. 2004. The promise of a dna taxonomy. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 359:669—679.

Bleidorn C. 2017. Phylogenomics: An Introduction. Springer.

Bodt SD, Maere S, de Peer YV. 2005. Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*. 20:591–597.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 30:2114–2120.

Bolton B. 2003. Synopsis and classification of Formicidae. American Entomological Institute.

Bolton B, et al. (2 co-authors). 1994. Identification guide to the ant genera of the world. Harvard University Press.

Bonnet T, Leblois R, Rousset F, Crochet PA. 2017. A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *Evolution*. pp. n/a–n/a.

Bouckaert R, Heled J. 2014. Densitree 2: Seeing trees through the forest. *bioRxiv*. .

Bribiesca-Contreras G, Solís-Marín FA, Laguarda-Figueras A, Zaldívar-Riverón A. 2013. Identification of echinoderms (echinodermata) from an anchialine cave in cozumel island, mexico, using dna barcodes. *Molecular ecology resources*. 13:1137–1145.

Brito PH, Edwards SV. 2009. Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*. 135:439–455.

Bronstein JL. 1994. Our current understanding of mutualism. *The Quarterly Review of Biology*. 69:31–51.

Bronstein JL. 1998. The contribution of ant-plant protection studies to our understanding of mutualism. *Biotropica*. 30:150–161.

Bronstein JL, Alarcón R, Geber M. 2006. The evolution of plant-insect mutualisms. *New Phytologist*. 172:412–428.

Brower AV. 2006. Problems with dna barcodes for species delimitation: Ten species of *Astraptes fulgerator* reassessed (lepidoptera: Hesperiidae). *Systematics and Biodiversity*. 4:127–132.

Brower AVZ, DeSalle R, Vogler A. 1996. Gene trees, species trees, and systematics: A cladistic perspective. *Annual Review of Ecology and Systematics*. 27:423–450.

Brumfield RT, Edwards SV. 2007. Evolution into and out of the andes: A bayesian analysis of historical historical diversification in *Thamnophilus antshrikes*. *Evolution*. 61:346–367.

Bruna EM, Izzo TJ, Inouye BD, Uriarte M, Vasconcelos HL. 2011. Asymmetric dispersal and colonization success of amazonian plant-ants queens. *PLOS ONE*. 6:e22937.

Bruna EM, Vasconcelos HL, Heredia S. 2005. The effect of habitat fragmentation on communities of mutualists: Amazonian ants and their host plants. *Biological Conservation*. 124:209–216.

Burger WC. 1981. Why are there so many kinds of flowering plants? *BioScience*. 31:572–581.

C MS, L VH. 2009. Long-term persistence of a neotropical ant-plant population in the absence of obligate plant-ants. *Ecology*. 90:2375–2383.

Cabrera M, Jaffé K. 1994. A trophic mutualism between the myrmecophytic melastomataceae *Tococa guianensis* aublet and an *Azteca* ant species. *Ecotropicos*. 7:1—10.

Cahill JA, Green RE, Fulton TL, et al. (11 co-authors). 2013. Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. *PLOS Genetics*. 9:1–8.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. Blast+: architecture and applications. *BMC bioinformatics*. 10:421.

Campbell B, Steffen-Campbell J, Werren J. 1994. Phylogeny of the *Nasonia* species complex (hymenoptera: Pteromalidae) inferred from an internal transcribed spacer (its2) and 28s rdna sequences. *Insect molecular biology*. 2:225–237.

Cardoso DC, Cristiano MP, Barros LAC, Lopes DM, Pompolo SdG. 2012. *First cytogenetic characterization of a species of the arboreal ant genus Azteca forel, 1978 (dolichoderinae, formicidae)*. 6:107–114.

CBOL PWg, Hollingsworth PM, Forrest LL, et al. (11 co-authors). 2009. A dna barcode for land plants. *Proceedings of the National Academy of Sciences*. 106:12794–12797.

Cediel F, Shaw RP, Cceres C. 2003. The Circum-Gulf of Mexico and the Caribbean: Hydrocarbon habitats, basin formation, and plate tectonics, AAPG Special Volumes, chapter Tectonic assembly of the northern Andean block, pp. 815–848.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*. 10.

Chen C, Renner S. 2007. Flora of China, St. Louis: Science Press, Beijing & Missouri Botanical Garden Press, chapter Melastomataceae, pp. 360–399.

Chen S, Yao H, Han J, et al. (16 co-authors). 2010. Validation of the its2 region as a novel dna barcode for identifying medicinal plant species. *PLOS ONE*. 5:e8613.

Chenuil A, McKey DB. 1996. Molecular phylogenetic study of a myrmecophyte symbiosis: Did *Leonardoxa*-ant associations diversify via cospeciation? *Molecular Phylogenetics and Evolution*. 6:270–286.

Chiotis M, Jermiin LS, Crozier RH. 2000. A molecular framework for the phylogeny of the ant subfamily dolichoderinae. *Molecular Phylogenetics and Evolution*. 17:108–116.

Chomicki G, Renner SS. 2015. Phylogenetics and molecular clocks reveal the repeated evolution of ant-plants after the late miocene in africa and the early miocene in australasia and the neotropics. *New Phytologist*. 207:411—424.

Chomicki G, Renner SS. 2016. Evolutionary relationships and biogeography of the ant-epiphytic genus *Squamellaria* (rubiaceae: Psychotrieae) and their taxonomic implications. *PLOS ONE*. 11:1–24.

Chomicki G, Ward PS, Renner SS. 2015. Macroevolutionary assembly of ant/plant symbioses: *Pseudomyrmex* ants and their ant-housing plants in the neotropics. *Proceedings of the Royal Society of London B: Biological Sciences*. 282.

Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N. 2012. Why assembling plant genome sequences is so challenging. *Biology*. 1:439–459.

Clausing G. 1998. Observations on ant-plant interactions in *Pachycentria* and other genera of the dissochaeteae (melastomataceae) in sabah and sarawak. *Flora*. 193:361–368.

Clausing G. 2000. Revision of *Pachycentria* (melastomataceae). *Blumea*. 45:341–375.

Clausing G, Renner SS. 2001. Evolution of growth form in epiphytic dissochaeteae (melastomataceae). *Organisms Diversity & Evolution*. 1:45–60.

Cognato AI. 2006. Standard percent dna sequence difference for insects does not predict species boundaries. *Journal of Economic Entomology*. 99:1037–1045.

Cogniaux A. 1891. Monographie phanerogamarum, volume 7, chapter Melastomataceae, pp. 1–1256.

Coleman AW, Vacquier VD. 2002. Exploring the phylogenetic utility of its sequences for animals: A test case for abalone (*Haliotis*). *Journal of Molecular Evolution*. 54:246–257.

Collins RA, Armstrong KF, Meier R, Yi Y, Brown SDJ, Cruickshank RH, Keeling S, Johnston C. 2012. Barcoding and border biosecurity: Identifying cyprinid fishes in the aquarium trade. *PLOS ONE*. 7:1–13.

Collins RA, Cruickshank RH. 2013. The seven deadly sins of dna barcoding. *Molecular Ecology Resources*. 13:969—975.

Cornils A, Held C. 2014. Evidence of cryptic and pseudocryptic speciation in the *Paracalanus parvus* species complex (crustacea, copepoda, calanoida). *Frontiers in Zoology*. 11:19.

Crane PR, Friis EM, Pedersen KR. 1995. The origin and early diversification of angiosperms. *Nature*. 374:27.

Crisan A, Munzner T, Gardy JL. 2018. Adjutant: an r-based tool to support topic discovery for systematic and literature reviews. *bioRxiv*. .

Dale H Clayton JM. 1997. Host-parasite evolution: General principles and avian models, Oxford Univerity Press, England., chapter Collection and quantification of arthropod parasites of birds, pp. 419–440.

Dalla Torre KW. 1894. Catalogus Hymenopterorum hucusque descriptorum systematicus et synonymicus, volume 1. G. Engelmann.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jmodeltest 2: more models, new heuristics and parallel computing. *Nat Meth*. 9:772–772.

Dáttilo WFC, Izzo TJ, Inouye BD, Vasconcelos HL, Bruna EM. 2009. Recognition of host plant volatiles by *Pheidole minutula* mayr (myrmicinae), an amazonian ant-plant specialist: Host-plant selection by a plant-ant. *Biotropica*. 41:642—646.

Davidson D, Fisher B. 1991. Symbiosis of ants with *Cecropia* as a function of light regime. *Huxley, C, R,, Cutler, D, F ed (s). Ant-plant interactions. Oxford Univ. Press: Oxford, etc*. pp. 289–309.

Davidson DM Diane West. 1993. The evolutionary ecology of symbiotic ant-plat relationships. *Journal of Hymenoptera Research*. 2:13–83.

Davidson DW, McKey D. 1993. *Ant-plant symbioses: Stalking the chuyachaqui*. 8:326–332.

Davidson DW, Snelling RR, Longino JT. 1989. Competition among ants for myrmecophytes and the significance of plant trichomes. *Biotropica*. 21:64–73.

Davidson R, Vachaspati P, Mirarab S, Warnow T. 2015. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC genomics*. 16:S1.

Davies SJ, Lum SKY, Chan R, Wang LK. 2001. Evolution of myrmecophytism in western malesian *Macaranga* (euphorbiaceae). *Evolution*. 55:1542–1559.

Davison J, Ho SY, Bray SC, et al. (12 co-authors). 2011. Late-quaternary biogeographic scenarios for the brown bear (ursus arctos), a wild mammal model species. *Quaternary Science Reviews*. 30:418–430.

De Bary A. 1879. Die erscheinung der symbiose. Verlag von Karl J. Trübner.

de Candolle A. 1828. Melastomataceae. *Prodromus Systematis Naturalis Regny Vegetabilis*. 3:99–202.

de González Juana C. 1980. Geologia de venezuela y de sus cuencas petroliferas. 2 vols. *Ediciones Foninves*. 1031.

de Vienne, D M, Refrégier G, López-Villavicencio M, Tellier A, Hood ME, Giraud T. 2013. Cospeciation vs host-shift speciation: methods for testing, evidence from natural associations and relation to coevolution. *New Phytol*. 198:347–385.

Dean MD, Ballard KJ, Glass A, William J, Ballard O. 2003. Influence of two wolbachia strains on population structure of east african *Drosophila simulans*. *Genetics*. 165:1959–1969.

Debout GD, Ventelon-Debout M, Emerson BC, Yu DW. 2007. Pcr primers for polymorphic microsatellite loci in the plant-ant *Azteca ulei cordiae* (formicidae: Dolichoderinae). *Molecular Ecology Resources*. 7:607–609.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*. 24:332–340.

DeHeer CJ, Tscbinkel WR. 1998. The success of alternative reproductive tactics in monogyne populations of the ant *Solenopsis invicta*: significance for transitions in social organization. *Behavioral Ecology*. 9:130–135.

Dejean A, Grangier J, Leroy C, Orivel J. 2008. Predation and aggressiveness in host plant protection: a generalization using ants from the genus *Azteca*. *Naturwissenschaften*. 96:57—63.

Dev SA, Shenoy M, Borges RM. 2010. Genetic and clonal diversity of the endemic ant-plant *Humboldtia brunonis* (fabaceae) in the western ghats of india. *Journal of Biosciences*. 35:267–279.

Doležel J, Greilhuber J, Suda J. 2007. Estimation of nuclear dna content in plants using flow cytometry. *Nature Protocols*. 2:2233.

Dominik R Laetsch MB. 2017. Blobtools: Interrogation of gneome assemblies [version 1; referees: 1 approved with reservations]. *F1000Research*. 6:1287.

Dong W, Xu C, Li C, Sun J, Zuo Y, Shi S, Cheng T, Guo J, Zhou S. 2015. ycf1, the most promising plastid dna barcode of land plants. *Scientific Reports*. 5:8348 EP.

Donoghue MJ, Moore BR. 2003. Toward an integrative historical biogeography. *Integrative and Comparative Biology*. 43:261–270.

Dormann CF, Gruber B, Fründ J. 2008. Introducing the bipartite package: analysing ecological networks. *interaction*. 1:0–2413793.

Douglas AE. 2010. The symbiotic habit. Princeton University Press.

Doyle JJ. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. *Systematic Botany*. 17:144–163.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol*. 4.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with beauti and the beast 1.7. *Molecular biology and evolution*. 29:1969–1973.

DuBay SG, Witt CC. 2014. Differential high-altitude adaptation and restricted gene flow across a mid-elevation hybrid zone in andean tit-tyrant flycatchers. *Molecular Ecology*. 23:3551–3565.

Dupuis JR, Roe AD, Sperling FAH. 2012. Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology*. 21:4422–4436.

Edgar RC. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 32:1792–1797.

Edwards DP, Hassall M, Sutherland WJ, Yu DW. 2006. Assembling a mutualism: ant symbionts locate their host plants by detecting volatile chemicals. *Insectes Sociaux*. 53:172–176.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution*. 63:1—19.

Ehlers TA, Poulsen CJ. 2009. Influence of andean uplift on climate and paleoaltimetry estimates. *Earth and Planetary Science Letters*. 281:238–248.

Ehrlich PR, Raven PH. 1964. Butterflies and plants: A study in coevolution. *Evolution*. 18:586–608.

El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research*. 24:831–838.

Elias M, Hill RI, Willmott KR, Dasmahapatra KK, Brower AV, Mallet J, Jiggins CD. 2007. Limited performance of dna barcoding in a diverse community of tropical butterflies. *Proceedings of the Royal Society of London B: Biological Sciences*. 274:2881–2889.

Elliott TA, Gregory TR. 2015. Whats in a genome? the c-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 370.

Ellstrand NC, Schierenbeck KA. 2000. Hybridization as a stimulus for the evolution of invasiveness in plants? *Proceedings of the National Academy of Sciences*. 97:7043–7050.

Emery C. 1893. *Studio monografico sul genere Azteca forel.* .

Emery C. 1913. Études sur les myrmicinae. *Annales.* .

Eyer P, Leniaud L, Tinaut A, Aron S. 2016. Combined hybridization and mitochondrial capture shape complex phylogeographic patterns in hybridogenetic *Cataglyphis* desert ants. *Molecular Phylogenetics and Evolution*. 105:251–262.

Fayle TM, Edwards DP, Turner EC, Dumbrell AJ, Eggleton P, Foster WA. 2011. Public goods, public services and by-product mutualism in an ant-fern symbiosis. *Oikos*. 121:1279–1286.

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH. 2008. Multiple multilocus dna barcodes from the plastid genome discriminate plant species equally well. *PLOS ONE*. 3:e2802.

Fazekas AJ, Kesanakurti PR, Burgess KS, Percy DM, Graham SW, Barrett SCH, Newmaster SG, Hajibabaei M, Husband BC. 2009. Are plant species inherently harder to discriminate than animal species using dna barcoding markers? *Molecular Ecology Resources*. 9:130–139.

Feldhaar H, Fiala B, Gadau J, Mohamed M, Maschwitz U. 2003. Molecular phylogeny of *Crematogaster* subgenus decacrema ants (hymenoptera: Formicidae) and the colonization of *Macaranga* (euphorbiaceae) trees. *Molecular Phylogenetics and Evolution*. 27:441–452.

Feldhaar H, Foitzik S, Heinze J. 2008. Lifelong commitment to the wrong partner: hybridization in ants. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 363:2891–2899.

Fiala B, Jakob A, Maschwitz U, Linsenmair KE. 1999. Diversity, evolutionary specialization and geographic distribution of a mutualistic ant-plant complex: *Macaranga* and *Crematogaster* in south east asia. *Biological journal of the Linnean Society*. 66:305–331.

Fisher B. 2010. Ant ecology, Oxford University Press, chapter Biogeography, pp. 18–37.

Florea L, Souvorov A, Kalbfleisch TS, Salzberg SL. 2011. Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. *PLoS One*. 6:e21400.

Fonseca CR, Ganade G. 1996. Asymmetries, compartments and null interactions in an amazonian ant-plant community. *Journal of Animal Ecology*. 65:339–347.

Forel A. 1878. Études myrmécologiques en 1878. *Bull. Soc. Vaud. Sci. Nat.* 15:337–392.

Forel A, Ogden CK. 1928. Social World Of The Ants Compared With That Of Man. GP Putnam'S Sons, Ltd.; London.

Fowler H. 1993. Relative representation of *Pheidole* (hymenoptera: Formicidae) in local ground ant assemblages of the americas. In: Anales de Biologia: Seccion Biologia Animal; Biologia Vegetal; Biologia Ambiental. volume 19, pp. 29–37.

Fox LR. 1988. Diffuse coevolution within complex communities. *Ecology*. 69:906–907.

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. 2008. Many or most genes in *Arabidopsis* transposed after the origin of the order brassicales. *Genome Research*. 18:1924–1937.

Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C. 2012. Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*. 27:480–488.

Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial dna. *Annual Review of Ecology, Evolution, and Systematics*. 34:397–423.

Futuyma DJ. 2009. Chapter 51 - coevolution. In: Resh VH, Cardé RT, editors, Encyclopedia of Insects (Second Edition), San Diego: Academic Press, pp. 175–179. Second edition edition.

Gernhard T. 2008. *The conditioned reconstructed process*. 253:769–778.

Goitia W, Jaffe K. 2009. Ant-plant associations in different forests in venezuela. *Neotropical Entomology*. 38:7–31.

Goldenberg R, Almeda F, Caddah MK, Martins AB, Meirelles J, Michelangeli FA, Weiss M. 2013. Nomenclator botanicus for the neotropical genus *Miconia* (melastomataceae: Miconieae). *Phytotaxa*. 106:1–171.

Goldenberg R, Penneys DS, Almeda F, Judd WS, Michelangeli FA. 2008. Phylogeny of *Miconia* (melastomataceae): Patterns of stamen diversification in a megadiverse neotropical genus. *International Journal of Plant Sciences*. 169:963—979.

Gómez-Acevedo S, Rico-Arce L, Delgado-Salinas A, Magallón S, Eguiarte LE. 2010. Neotropical mutualism between *Acacia* and *Pseudomyrmex*: Phylogeny and divergence times. *Molecular Phylogenetics and Evolution*. 56:393–408.

Gómez-Zurita J, Juan C, Petitpierre E. 2000. Sequence, secondary structure and phylogenetic analyses of the ribosomal internal transcribed spacer 2 (its2) in the *Timarcha* leaf beetles (coleoptera: Chrysomelidae). *Insect Molecular Biology*. 9:591–604.

Gore MA, Chia JM, Elshire RJ, et al. (11 co-authors). 2009. A first-generation haplotype map of maize. *Science*. 326:1115–1117.

Grant BR, Grant PR. 1996. High survival of darwin's finch hybrids: effects of beak morphology and diets. *Ecology*. 77:500–509.

Gregory-Wodzicki KM. 2000. Uplift history of the central and northern andes: a review. *Geological Society of America Bulletin*. 112:1091—1105.

Greilhuber J, Borsch T, Müller K, Worberg A, Porembski S, Barthlott W. 2006. Smallest angiosperm genomes found in lentibulariaceae, with chromosomes of bacterial size. *Plant Biology*. 8:770–777.

Grimaldi D, Engel MS. 2005. Evolution of the Insects. Cambridge University Press.

Guerrero RJ, Delabie JH, Dejean A. 2010. Taxonomic contribution to the aurita group of the ant genus *Azteca* (formicidae: Dolichoderinae). *J Hymenopt Res*. 19:51–65.

Hafner MS, Nadler SA. 1990. Cospeciation in host-parasite assemblages: Comparative analysis of rates of evolution and timing of cospeciation events. *Systematic Biology*. 39:192–204.

Hahn C, Bachmann L, Chevreux B. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - a baiting and iterative mapping approach. *Nucleic Acids Research*. 41:e129.

Hailer F, Kutschera VE, Hallström BM, Klassert D, Fain SR, Leonard JA, Arnason U, Janke A. 2012. Nuclear genomic sequences reveal that polar bears are an old and distinct bear lineage. *Science*. 336:344–347.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA. 2007. *Dna barcoding: how it complements taxonomy, molecular phylogenetics and population genetics*. 23:167–172.

Hamilton CA, Hendrixson BE, Brewer MS, Bond JE. 2014. An evaluation of sampling effects on multiple dna barcoding methods leads to an integrative approach for delimiting species: a case study of the north american tarantula genus aphonopelma

(araneae, mygalomorphae, theraphosidae). *Molecular Phylogenetics and Evolution*. 71:79—93.

Hearn J, Stone GN, Bunnefeld L, Nicholls JA, Barton NH, Lohse K. 2014. Likelihood-based inference of population history from low-coverage *de novo* genome assemblies. *Molecular Ecology*. 23:198—211.

Hebert PD, Ratnasingham S, de Waard JR. 2003a. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*. 270:S96–S99.

Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003b. Biological identifications through dna barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*. 270:313—321.

Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W. 2004a. Ten species in one: Dna barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*. 101:14812–14817.

Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM. 2004b. Identification of birds through dna barcodes. *PLOS Biology*. 2.

Heil M, González-Teuber M, Clement LW, Kautz S, Verhaagh M, Bueno JCS. 2009. Divergent investment strategies of *Acacia* myrmecophytes and the coexistence of mutualists and exploiters. *Proceedings of the National Academy of Sciences*. 106:18091–18096.

Heil M, McKey D. 2003. Protective ant-plant interactions as model systems in ecological and evolutionary research. *Annual Review of Ecology, Evolution, and Systematics*. 34:425–553.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*. 8:289.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*. 27:570.

Heled J, Drummond AJ. 2015. Calibrated birth–death phylogenetic time-tree priors for bayesian inference. *Systematic Biology*. 64:369–383.

Helms JA, Kaspari M. 2015. Reproduction-dispersal tradeoffs in ant queens. *Insectes Sociaux*. 62:171–181.

Hembry DH, Yoder JB, Goodman KR. 2014. Coevolution and the diversification of life. *The American Naturalist*. 184:425–438.

Herre EA, Knowlton N, Mueller UG, Rehner SA. 1999. *The evolution of mutualisms: exploring the paths between conflict and cooperation*. 14:49–53.

Ho SYW, Lo N. 2013. The insect molecular clock. *Australian Journal of Entomology*. 52:101–105.

Hölldobler B, Engel-Siegel H. 1984. On the metapleural gland of ants. *Psyche*. 91:201–224.

Hölldobler B, Wilson EO. 1990. The ants. Harvard University Press.

Hollingsworth PM, Graham SW, Little DP. 2011. Choosing and using a plant dna barcode. *PLOS ONE*. 6:1–13.

Hoorn C. 1993. Marine incursions and the influence of andean tectonics on the miocene depositional history of northwestern amazonia: results of a palynostratigraphic study. *Palaeogeography, Palaeoclimatology, Palaeoecology*. 105:267–309.

Hoorn C, Guerrero J, Sarmiento GA, Lorente MA. 1995. Andean tectonics as a cause for changing drainage patterns in miocene northern south america. *Geology*. 23:237–240.

Hoorn C, Wesselingh FP, ter Steege H, et al. (18 co-authors). 2010. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science*. 330:927–931.

Hughes C, Eastwood R. 2006. Island radiation on a continental scale: Exceptional rates of plant diversification after uplift of the andes. *Proceedings of the National Academy of Sciences*. 103:10334–10339.

Hung YT, Chen CA, Wu WJ, Lin CC, Shih CJ. 2004. Phylogenetic utility of the ribosomal internal transcribed spacer 2 in *Strumigenys* spp. (hymenoptera: Formicidae). *Molecular Phylogenetics and Evolution*. 32:407–415.

Ilinsky Y. 2013. Coevolution of *Drosophila melanogaster* mtdna and *Wolbachia* genotypes. *PLOS ONE*. 8:e54373.

Irwin DE. 2002. Phylogeographic breaks without geographic barriers to gene flow. *Evolution*. 56:2383–2394.

Jansen G, Savolainen R, Vepsäläinen K. 2009. Dna barcoding as a heuristic tool for classifying undescribed nearctic *Myrmica* ants (hymenoptera: Formicidae). *Zoologica Scripta*. 38:527–536.

Janz N. 2011. Ehrlich and raven revisited: Mechanisms underlying codiversification of plants and enemies. *Annual Review of Ecology, Evolution, and Systematics*. 42:71–89.

Janzen DH. 1966. Coevolution of mutualism between ants and acacias in central america. *Evolution*. 20:249–275.

Janzen DH. 1980. When is it coevolution. *Evolution*. 34:611–612.

Janzen DH. 1985. The natural history of mutualisms.

Janzen DH, Hallwachs W. 2011. Joining inventory by parataxonomists with dna barcoding of a large complex tropical conserved wildland in northwestern costa rica. *PLOS ONE*. 6:1–13.

Jaramillo C, Hoorn C, Silva SAF, Leite F, Herrera F, Quiroz L, Dino R, Antonioli L. 2009. The Origin of the Modern Amazon Rainforest: Implications of the Palynological and Palaeobotanical Record, Wiley-Blackwell Publishing Ltd., pp. 317–334.

Jarvis DE, Ho YS, Lightfoot DJ, et al. (33 co-authors). 2017. The genome of *Chenopodium quinoa*. *Nature*. 542:307.

Jarvis ED, Mirarab S, Aberer AJ, et al. (105 co-authors). 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*. 346:1320–1331.

Jeanson ML, Labat JN, Little DP. 2011. Dna barcoding: a new tool for palm taxonomists? *Annals of botany*. 108:1445–1451.

Jiao Y, Wickett NJ, Ayyampalayam S, et al. (17 co-authors). 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature*. 473:97–100.

Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*. 174:E54–E70. PMID: 19519219.

Jones G, Aydin Z, Oxelman B. 2014. Dissect: an assignment-free bayesian discovery method for species delimitation under the multispecies coalescent. *Bioinformatics*. 31:991–998.

Jones M, Ghoorah A, Blaxter M. 2011. jmotu and taxonerator: Turning dna barcode sequences into annotated operational taxonomic units. *PLoS ONE*. 6:1–10.

Jordano P. 2000. Seeds: the ecology of regeneration in plant communities, volume 2, chapter Fruits and frugivory, pp. 125–166.

Jousselin E, Desdevises Y, Coeur d'Acier A. 2009. Fine-scale cospeciation between *Brachycaudus* and *Buchnera aphidicola*: bacterial genome helps define species and evolutionary relationships in aphids. *Proceedings of the Royal Society of London B: Biological Sciences*. 276:187–196.

Julliard R, Clavel J, Devictor V, Jiguet F, Couvet D. 2006. Spatial segregation of specialists and generalists in bird communities. *Ecology Letters*. 9:1237—1244.

Jürgens A, Feldhaar H, Feldmeyer B, Fiala B. 2006. *Chemical composition of leaf volatiles in Macaranga species (euphorbiaceae) and their potential role as olfactory cues in host-localization of foundress queens of specific ant partners*. 34:97–113.

Kajitani R, Toshimoto K, Noguchi H, et al. (11 co-authors). 2014. Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*. 24:1384—1395.

Karsten S, Boytd B, C WJ, Emma N, G GM, John B, W EG, A TJ. 2008. When rare species become endangered: cryptic speciation in myrmecophilous hoverflies. *Biological Journal of the Linnean Society*. 75:291–300.

Kearse M, Moir R, Wilson A, et al. (14 co-authors). 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 28:1647.

Keller L. 1998. Queen lifespan and colony characteristics in ants and termites. *Insectes Sociaux*. 45:235–246.

Kiester AR, Lande R, Schemske DW. 1984. Models of coevolution and speciation in plants and their pollinators. *The American Naturalist*. 124:220–243.

Kingman JFC. 1982. *The coalescent*. 13:235–248.

Kobert K, Salichos L, Rokas A, Stamatakis A. 2016. Computing the internode certainty and related measures from partial gene trees. *Molecular biology and evolution*. 33:1606–1617.

Koptur S, Rico-Gray V, Palacios-Rios M. 1998. Ant protection of the nectaried fern polypodium plebeium in central mexico. *American Journal of Botany*. 85:736–736.

Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker AA, Blaxter M. 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proceedings of the National Academy of Sciences*. 113:5053–5058.

Kress WJ, Erickson DL. 2007. A two-locus global dna barcode for land plants: The coding rbcl gene complements the non-coding trnh-psba spacer region. *PLOS ONE*. 2:1–10.

Kress WJ, Erickson DL. 2008. Dna barcodes: Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences of the United States of America*. 105:2761–2762.

Kriebel R, Michelangeli FA, Kelly LM. 2015. Discovery of unusual anatomical and continuous characters in the evolutionary history of *Conostegia* (miconieae: Melastomataceae). *Molecular Phylogenetics and Evolution*. 82:289–313.

Kubatko LS, Carstens BC, Knowles LL. 2009. Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*. 25:971–973.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*. 56:17–24.

Kumar S, Banks TW, Cloutier S. 2012. Snp discovery through next-generation sequencing and its applications. *International journal of plant genomics*. 2012.

Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated gc-coverage plots. *Frontiers in Genetics*. 4:237.

Kumar S, Stecher G, Tamura K. 2016. Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular biology and evolution*. 33:1870–1874.

Kutschera VE, Bidon T, Hailer F, Rodi JL, Fain SR, Janke A. 2014. Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Molecular Biology and Evolution*. 31:2004–2017.

Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V. 2008. Dna barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences*. 105:2923–2928.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2016. Partitionfinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*. p. msw260.

Lapola DM, Bruna EM, Vasconcelos HL. 2003. Contrasting responses to induction cues by ants inhabiting *Maieta guianensis* (melastomataceae). *Biotropica*. 35:295–300.

Laurent S, Pfeifer SP, Settles ML, Hunter SS, Hardwick KM, Ormond L, Sousa VC, Jensen JD, Rosenblum EB. 2016. The population genomics of rapid adaptation: disentangling signatures of selection and demography in white sands lizards. *Molecular Ecology*. 25:306–323.

Leaché AD, Banbury BL, Linkem CW, de Oca ANM. 2016. Phylogenomics of a rapid radiation: is chromosomal evolution linked to increased diversification in north american spiny lizards (genus *Sceloporus*)? *BMC Evolutionary Biology*. 16:63.

Leigh EG. 2010. The evolution of mutualism. *Journal of Evolutionary Biology*. 23:2507–2528.

Leotard G, Saltmarsh A, Kjellberg F, Mckey D. 2008. Mutualism, hybrid inviability and speciation in a tropical ant-plant. *Journal of Evolutionary Biology*. 21:1133–1143.

Li H, Durbin R. 2009. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*. 25:1754–60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and samtools. *Bioinformatics (Oxford, England)*. 25:2078–9.

Li JH, jian Liu Z, Salazar GA, Bernhardt P, Perner H, Tomohisa Y, hua Jin X, wen Chung S, bo Luo Y. 2011. Molecular phylogeny of *Cypripedium* (orchidaceae: Cypripedioideae) inferred from multiple nuclear and chloroplast regions. *Molecular Phylogenetics and Evolution*. 61:308–320.

Li Y, Zhou X, Feng G, Hu H, Niu L, Hebert PDN, Huang D. 2010. Coi and its2 sequences delimit species, reveal cryptic taxa and host specificity of fig-associated *Sycophila* (hymenoptera, eurytomidae). *Molecular Ecology Resources*. 10:31–40.

Librado P, Rozas J. 2009. Dnasp v5: a software for comprehensive analysis of dna polymorphism data. *Bioinformatics*. 25:1451.

Linares MC, Soto-Calderón ID, Lees DC, Anthony NM. 2009. High mitochondrial diversity in geographically widespread butterflies of madagascar: A test of the dna barcoding approach. *Molecular Phylogenetics and Evolution*. 50:485–495.

Linder CR, Rieseberg LH. 2004. Reconstructing patterns of reticulate evolution in plants. *American journal of botany*. 91:1700–1708.

Linder HP. 2008. Plant species radiations: where, when, why? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 363:3097–3105.

Lindqvist C, Schuster SC, Sun Y, et al. (14 co-authors). 2010. Complete mitochondrial genome of a pleistocene jawbone unveils the origin of polar bear. *Proceedings of the National Academy of Sciences*. 107:5053–5057.

Lisch D. 2013. How important are transposons for plant evolution? *Nat Rev Genet*. 14:49–61.

Little DP. 2011. Dna barcode sequence identification incorporating taxonomic hierarchy and within taxon variability. *PLOS ONE*. 6:1–1.

Liu L, Wu S, Yu L. 2015a. Coalescent methods for estimating species trees from phylogenomic data. *Journal of systematics and evolution*. 53:380–390.

Liu L, Xi Z, Wu S, Davis CC, Edwards SV. 2015b. Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*. 1360:36–53.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*. 10:302.

Lohse K. 2009. Can mtdna barcodes be used to delimit species? a response to pons et al. (2006). *Systematic Biology*. 58:439–442.

Londoño GA, Chappell MA, Jankowski JE, Robinson SK. 2017. Do thermoregulatory costs limit altitude distributions of andean forest birds? *Functional Ecology*. 31:204–215.

Longino J. 1991a. Taxonomy of the *Cecropia*-inhabiting *Azteca* ants. *Journal of Natural History*. 25:1571–1602.

Longino JT. 1986. Ants provide substrate for epiphytes. *Selbyana*. pp. 100–103.

Longino JT. 1989. Geographic variation and community structure in an ant-plant mutualism: *Azteca* and *Cecropia* in costa rica. *Biotropica*. 21:126–132.

Longino JT. 1991b. Ant-plant interactions. Oxford University Press, 19, pp. 271–288.

Longino JT. 1991c. *Azteca ants in Cecropia trees: taxonomy, colony structure, and behaviour*. .

Longino JT. 1996. Taxonomic characterization of some live-stem inhabiting *Azteca* (hymenoptera: Formicidae) in costa rica, with special reference to the ants of *Cordia* (boraginaceae) and *Triplaris* (polygonaceae). *Journal of Hymenoptera Research*. 5:131–156.

Longino JT. 2007. A taxonomic review of the genus *Azteca* (Hymenoptera: Formicidae) in Costa Rica and a global revision of the aurita group. Magnolia Press.

Longino JT, Coddington J, Colwell RK. 2002. The ant fauna of a tropical rain forest: Estimating species richness three different ways. *Ecology*. 83:689–702.

Machado CA, Robbins N, Gilbert MTP, Herre EA. 2005. Critical review of host specificity and its coevolutionary implications in the fig/fig-wasp mutualism. *Proceedings of the National Academy of Sciences*. 102:6558–6565.

MacKay WP, Vinson SB. 1989. *A guide to species identification of new world ants (hymenoptera: Formicidae)*. .

Maddison W, Knowles L. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*. 55:21—30.

Malé PJG, Leroy C, Humblot P, Dejean A, Quilichini A, Orivel J. 2016. Limited gene dispersal and spatial genetic structure as stabilizing factors in an ant-plant mutualism. *Journal of Evolutionary Biology*. 29:2519–2529.

Mallet J. 2007. Hybrid speciation. *Nature*. 446:279–283.

Mallet J. 2009. Rapid speciation, hybridization and adaptive radiation in the heliconius melpomene group. *Speciation and patterns of diversity*. pp. 177–194.

Martin CV, Little DP, Goldenberg R, Michelangeli FA. 2008. A phylogenetic evaluation of *Leandra* (miconieae, melastomataceae): a polyphyletic genus where the seeds tell the story, not the petals. *Cladistics*. 24:315–327.

Martins J, Solomon ES, Mikheyev AS, Mueller UG, Ortiz A, Bacci M. 2007. Nuclear mitochondrial-like sequences in ants: evidence from *Atta cephalotes* (formicidae: Attini). *Insect Molecular Biology*. 16:777–784.

Martius C. 1832. *Nova genera et species plantarum*. .

McFarquhar AM, Robertson FW. 1963. The lack of evidence for co-adaptation in crosses between geographical races of *Drosophila subobscura* coll. *Genetical Research*. 4:104–131.

McNaughton S, Wolf L. 1973. General ecology. holt, rinehard and winston. *Inc., New York*. .

Meer RKV, Lofgren CS, Alvarez FM. 1985. Biochemical evidence for hybridization in fire ants. *The Florida Entomologist*. 68:501–506.

Meier R, Shiyang K, Vaidya G, Ng PKL. 2006. Dna barcoding and taxonomy in diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*. 55:715.

Meier R, Zhang G, Ali F, Zamudio K. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the barcoding gap and leads to misidentification. *Systematic Biology*. 57:809–813.

Meiklejohn KA, Faircloth BC, Glenn TC, Kimball RT, Braun EL. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. *Systematic biology*. p. syw014.

Mendoza ÁM, Torres MF, Paz A, Trujillo-Arias N, López-Alvarez D, Sierra S, Forero F, Gonzalez MA. 2016. Cryptic diversity revealed by dna barcoding in colombian illegally traded bird species. *Molecular ecology resources*. 16:862–873.

Meyer CP, Paulay G. 2005. Dna barcoding: Error rates based on comprehensive sampling. *PLOS Biology*. 3.

Meyers LA, Levin DA. 2006. On the abundance of polyploids in flowering plants. *Evolution*. 60:1198–1206.

Michelangeli FA. 2000. A cladistic analysis of the genus *Tococa* (melastomataceae) based on morphological data. *Systematic Botany*. 25:211.

Michelangeli FA. 2003. Ant protection against herbivory in three species of *Tococa* (melastomataceae) occupying different environments. *Biotropica*. 35:181–188.

Michelangeli FA. 2005. *Tococa* (melastomataceae). *Flora Neotropica*. pp. 1–114.

Michelangeli FA. 2010a. Neotropical myrmecophilous melastomataceae: an annotated list and key. *Proceedings of the California Academy of Sciences*. 61:409—449.

Michelangeli FA. 2010b. *Tococa guianensis*.

Michelangeli FA, Almeda F, Alvear M, Bécquer ER, Burke J, Caddah MK, Goldenberg R, Ionta GM, Judd WS, Majure LC. 2016. Proposal to conserve *Miconia*, nom. cons. against the additional names *Maieta* and *Tococa* (melastomataceae: Miconieae). *Taxon*. 65:892–893.

Michelangeli FA, Penneys DS, Giza J, Soltis D, Hils MH, Skean JD. 2004. A preliminary phylogeny of the tribe miconieae (melastomataceae) based on nrits sequence data and its implications on inflorescence position. *Taxon*. 53:279–279.

Michelangeli FA, s Judd W, Penneys DS, Skean JD, Bécquer-Granados ER, Goldenberg R, Martin CV. 2008. Multiple events of dispersal and radiation of the tribe miconieae (melastomataceae) in the caribbean. *The Botanical Review*. 74:53—77.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*. 30:i541–i548.

Mitchell A. 2015. Collecting in collections: a pcr strategy and primer set for dna barcoding of decades-old dried museum specimens. *Molecular ecology resources*. 15:1102–1111.

Moller M, Cronk Q. 1997. Origin and relationships of *Saintpaulia* (gesneriaceae) based on ribosomal dna internal transcribed spacer (its) sequences. *American Journal of Botany*. 84:956–956.

Morawetz W, Henzl M, Wallnöfer B. 1992. Tree killing by herbicide producing ants for the establishment of pure *Tococa occidentalis* populations in the peruvian amazon. *Biodiversity & Conservation*. 1:19–33.

Moreau CS, Bell CD. 2011. Fossil cross-validation of the dated ant phylogeny (hymenoptera: Formicidae). *Entomologica Americana*. 117:127—133.

Moreau CS, Bell CD. 2013. Testing the museum versus cradle tropical biological diversity hypothesis: Phylogeny, diversificatio, and ancestral biogeographic range evolution of the ants. *Evolution*. 67:2240—2257.

Moreau CS, Bell CD, Vila R, Archibald SB, Pierce NE. 2006. Phylogeny of the ants: Diversification in the age of angiosperms. *Science*. 312:101—104.

Moritz C, Fujita MK, Rosauer D, et al. (13 co-authors). 2016. Multilocus phylogeography reveals nested endemism in a gecko across the monsoonal tropics of australia. *Molecular Ecology*. 25:1354–1366.

Morley RJ, Dick CW. 2003. Missing fossils, molecular clocks, and the origin of the melastomataceae. *American Journal of Botany*. 90:1638–1644.

Mueller UG. 2002. Ant versus fungus versus mutualism: ant-cultivar conflict and the deconstruction of the attine ant-fungus symbiosis. *The American naturalist*. 160 Suppl 4:S67–98.

Murrell DJ, Travis JMJ, Dytham C. 2002. The evolution of dispersal distance in spatially-structured populations. *Oikos*. 97:229–236.

Myburg AA, Grattapaglia D, Tuskan GA, et al. (80 co-authors). 2014. The genome of *Eucalyptus grandis*. *Nature*. 510:356–362.

Nakashima S, Sarath E, Okada H, Ezaki K, Darnaedi D, Tsukaya H, Soejima A. 2016. Morphological and phylogenetic investigations for several cryptic ant-plants found in *Callicarpa* (lamiaceae) from borneo. *Journal of Plant Research*. 129:591–601.

Naudin C. 1851. Melastomacearum monographicae descriptiones. *Ann. Sci. Natl. Bot.* 3:83–246.

Neubig KM, Whitten WM, Carlsward BS, Blanco MA, Endara L, Williams NH, Moore M. 2009. Phylogenetic utility of ycf1 in orchids: a plastid gene more variable than matk. *Plant Systematics and Evolution*. 277:75–84.

Newmaster SG, Fazekas AJ, Ragupathy S. 2006. Dna barcoding in land plants: evaluation of rbcl in a multigene tiered approach. *Can. J. Bot.* 84:335–341.

Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J. 2008. Testing candidate plant barcode regions in the myristicaceae. *Molecular Ecology Resources*. 8:480–490.

Ngéndo RN, Osiemo ZB, Brandl R. 2013. Dna barcodes for species identification in the hyperdiverse ant genus pheidole (formicidae: Myrmicinae). *Journal of Insect Science*. 13:1–13.

Nicholls JA, Challis RJ, Mutun S, Stone GN. 2012. Mitochondrial barcodes are diagnostic of shared refugia but not species in hybridizing oak gallwasps. *Molecular Ecology*. 21:4051–4062.

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (leguminosae: Mimosoideae). *Frontiers in Plant Science*. 6:710.

Nurk S, Bankevich A, Antipov D, et al. (18 co-authors). 2013. Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 158–170.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaspades: a new versatile metagenomic assembler. *Genome Research*. 27:824–834.

Nwani CD, Becker S, Braid HE, Ude EF, Okogwu OI, Hanner R. 2011. Dna barcoding discriminates freshwater fishes from southeastern nigeria and provides river system-level phylogeographic resolution within some species. *Mitochondrial DNA*. 22:43–51.

Ocampo G, Michelangeli FA, Almeda F. 2014. Seed diversity in the tribe miconieae (melastomataceae): Taxonomic, systematic, and evolutionary implications. *PLoS ONE*. 9:e100561.

Okita I, Tsuchida K. 2016. Clonal reproduction with androgenesis and somatic recombination: the case of the ant *Cardiocondyla kagutsuchi*. *The Science of Nature*. 103:22.

O'Leary NA, Wright MW, Rodney BJ, et al. (55 co-authors). 2016. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*. 44:D733–D745.

Olmstead RG, Sweere JA. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the solanaceae. *Systematic Biology*. 43:467–481.

Olsson S, Seoane-Zonjic P, Bautista R, Claros GM, González-Martínez SC, Scotti I, Scotti-Saintagne C, Hardy OJ, Heuertz M. 2017. Development of genomic tools in a widespread tropical tree, *Symphonia globulifera* lf: a new low-coverage draft genome, snp and ssr markers. *Molecular ecology resources*. 17:614–630.

O'meara BC. 2009. New heuristic methods for joint species delimitation and species tree inference. *Systematic Biology*. 59:59–73.

Page R. 1993. Parasites, phylogeny and cospeciation. *International Journal for Parasitology*. 23:499–506. Special Issue: Proceedings of the Joint Conference of the Australian and New Zealand Societies for Parasitology, 1992.

Page RD. 2003. Tangled trees: phylogeny, cospeciation, and coevolution. University of Chicago Press.

Page RD, Charleston MA. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*. 7:231–240.

Pagès M, Calvignac S, Klein C, Paris M, Hughes S, Hänni C. 2008. Combined analysis of fourteen nuclear genes refines the ursidae phylogeny. *Molecular Phylogenetics and Evolution*. 47:73–83.

Paknia O, Bergmann T, Hadrys H. 2015. Some antswers: Application of a layered barcode approach to problems in ant taxonomy. *Mol Ecol Resour*. 15:1262–1274.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular biology and evolution*. 5:568–83.

Papadopoulou A, Anastasiou I, Vogler AP. 2010. Revisiting the insect mitochondrial molecular clock: The mid-aegean trench calibration. *Molecular Biology and Evolution*. 27:1659.

Paszkiewicz K, Studholme DJ. 2010. *De novo* assembly of short sequence reads. *Briefings in Bioinformatics*. 11:457–472.

Peccoud J, Piatscheck F, Yockteng R, et al. (12 co-authors). 2013. Multi-locus phylogenies of the genus *Barteria* (passifloraceae) portray complex patterns in the evolution of myrmecophytism. *Molecular Phylogenetics and Evolution*. 66:824–832.

Pellicer J, Fay MF, Leitch IJ. 2010. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society*. 164:10–15.

Pemberton RW. 1992. Fossil extrafloral nectaries, evidence for the ant-guard antiherbivore defense in an oligocene *Populus*. *American Journal of Botany*. pp. 1242–1246.

Peña C, Nylin S, Freitas AVL, Wahlberg N. 2010. Biogeographic history of the butterfly subtribe euptychiina (lepidoptera, nymphalidae, satyrinae). *Zoologica Scripta*. 39:243–258.

Penneys DS, Judd WS. 2011. Phylogenetics and morphology in the blakeeae (melastomataceae). *International Journal of Plant Sciences*. 172:78–106.

Pennington RT, Lavin M, Särkinen T, Lewis GP, Klitgaard BB, Hughes CE. 2010. Contrasting plant diversification histories within the andean biodiversity hotspot. *Proceedings of the National Academy of Sciences*. 107:13783–13787.

Percy DM. 2003. Radiation, diversity, and host-plant interactions among island and continental legume-feeding psyllids. *Evolution*. 57:2540–2556.

Percy DM, Page RDM, Cronk QCB. 2004. Plant-insect interactions: Double-dating associated insect and plant lineages reveals asynchronous radiations. *Systematic Biology*. 53:120–127.

Pérez-Escobar OA, Chomicki G, Condamine FL, Karremans AP, Bogarín D, Matzke NJ, Silvestro D, Antonelli A. 2017. Recent origin and rapid speciation of neotropical orchids in the worldś richest plant biodiversity hotspot. *New Phytologist*. 215:891–905. 2017-23782.

Petit RJ, Excoffier L. 2009. Gene flow and species delimitation. *Trends in Ecology & evolution*. 24:386–393.

Pirie MD, Chatrou LW, Mols JB, Erkens RHJ, Oosterhof J. 2006. 'andean-centred' genera in the short-branch clade of annonaceae: testing biogeographical hypotheses using phylogeny reconstruction and molecular dating. *Journal of Biogeography*. 33:31–46.

Plowman NS, Hood ASC, Moses J, Redmond C, Novotny V, Klimes P, Fayle TM. 2017. Network reorganization and breakdown of an ant-plant protection mutualism with elevation. *Proceedings of the Royal Society of London B: Biological Sciences*. 284.

Poisot T, Canard E, Mouquet N, Hochberg ME. 2012. A comparative study of ecological specialization estimators. *Methods in Ecology and Evolution*. 3:537–544.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. 2006. Sequence-based species delimitation for the dna taxonomy of undescribed insects. *Systematic Biology.* 55:595–609.

Pringle EG, Novo A, Ableson I, Barbehenn RV, Vannette RL. 2014. Plant-derived differences in the composition of aphid honeydew and their effects on colonies of aphid-tending ants. *Ecol Evol.* 4:4065–4079.

Pringle EG, RamíRez SR, Bonebrake TC, Gordon DM, Dirzo R. 2012. Diversification and phylogeographic structure in widespread *Azteca* plant-ants from the northern neotropics. *Molecular Ecology.* 21:3576—3592.

Puillandre N, Lambert A, Brouillet S, Achaz G. 2012. Abgd, automatic barcode gap discovery for primary species delimitation: Abgd, automatic barcode gap discovery. *Molecular Ecology.* 21:1864—1877.

Quek S, Davies SJ, Ashton PS, Itino T, Pierce NE. 2007. The geography of diversification in mutualistic ants: a geneś-eye view into the neogene history of sundaland rain forests. *Molecular Ecology.* 16:2045–2062.

Quek SP, Davies SJ, Itino T, Pierce NE, Pellmyr O. 2004. Codiversification in an ant-plant mutualism: Stem texture and the evolution of host use in *Crematogaster* (formicidea: Myrmecinae) inhabitants of *Macaranga* (euphorbiaceae). *Evolution.* 58:554–570.

Rabinowicz PD, Bennetzen JL. 2006. The maize genome as a model for efficient sequence analysis of large plant genomes. *Current opinion in plant biology.* 9:149–156.

Ramirez SR, Nieh JC, Quental TB, Roubik DW, Imperatriz-Fonseca VL, Pierce NE. 2010. A molecular phylogeny of the stingless bee genus *Melipona* (hymenoptera: Apidae). *Molecular phylogenetics and evolution*. 56:519–25.

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*. 164:1645–1656.

Rasko DA, Webster DR, Sahl JW, et al. (28 co-authors). 2011. Origins of the e. coli strain causing an outbreak of hemolytic–uremic syndrome in germany. *New England Journal of Medicine*. 365:709–717. PMID: 21793740.

Ratnasingham S, Hebert PDN. 2013. A dna-based registry for all animal species: The barcode index number (bin) system. *PLOS ONE*. 8:1–16.

Raven PH. 1977. A suggestion concerning the cretaceous rise to dominance of the angiosperms. *Evolution*. 31:451–452.

Razafimandimbison SG, Moog J, Lantz H, Maschwitz U, Bremer B. 2005. Re-assessment of monophyly, evolution of myrmecophytism, and rapid radiation in *Neonauclea* s.s. (rubiaceae). *Molecular Phylogenetics and Evolution*. 34:334–354.

Regal PJ. 1977. Ecology and evolution of flowering plant dominance. *Science*. 196:622–629.

Reginato M, Michelangeli FA. 2016. Primers for low-copy nuclear genes in the melastomataceae. *Applications in Plant Sciences*. 4:1500092.

Remfert J. 2012. Genetic Variation and Cluster Formation of the Ant *Azteca* in a Coffee Agroecosystem. Master's thesis, University of Michigan.

Ren BQ, Xiang XG, Chen ZD. 2010. Species identification of *Alnus* (betulaceae) using nrdna and cpdna genetic markers. *Molecular Ecology Resources*. 10:594–605.

Renner SS. 1989. A survey of reproductive biology in neotropical melastomataceae and memecylaceae. *Annals of the Missouri Botanical Garden*. 76:496–518.

Renner SS. 1993. Phylogeny and classification of the melastomataceae and memecylaceae. *Nordic Journal of Botany*. 13:519–540.

Renner SS, Clausing G, Meyer K. 2001. Historical biogeography of melastomataceae: the roles of tertiary migration and long-distance dispersal. *American Journal of Botany*. 88:1290–1300.

Renner SS, Ricklefs RE. 1998. Herbicidal activity of domatia-inhabiting ants in patches of *Tococa guianensis* and *Clidemia* heterophylla. *Biotropica*. 30:324–327.

Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM. 2001. Rapid diversification of a species-rich genus of neotropical rain forest trees. *Science*. 293:2242–2245.

Richardson JE, Whitlock BA, Meerow AW, Madriñán S. 2015. The age of chocolate: a diversification history of *Theobroma* and malvaceae. *Frontiers in Ecology and Evolution*. 3:120.

Rico-Gray V, Oliveira PS. 2007. The ecology and evolution of ant-plant interactions. University of Chicago Press.

Ridley M. 2003. Evolution. Oxford University Press.

Rieseberg LH, Raymond O, Rosenthal DM, Lai Z, Livingstone K, Nakazato T, Durphy JL, Schwarzbach AE, Donovan LA, Lexer C. 2003. Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*. 301:1211–1216.

Robinson JD, Bunnefeld L, Hearn J, Stone GN, Hickerson MJ. 2014. Abc inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology*. 23:4458–4471.

Roca AL, Georgiadis N, O'Brien SJ. 2005. Cytonuclear genomic dissociation in african elephant species. *Nature genetics*. 37.

Ronque MUV, Azevedo-Silva M, Mori GM, Souza AP, Oliveira PS. 2016. Three ways to distinguish species: using behavioural, ecological, and molecular data to tell apart two closely related ants, *Camponotus renggeri* and *Camponotus rufipes* (hymenoptera: Formicidae). *Zoological Journal of the Linnean Society*. 176:170—181.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*. 61:539.

Ross KG, Gotzek D, Ascunce MS, Shoemaker DD. 2009. Species delimitation: a case study in a problematic ant taxon. *Systematic biology*. 59:162–184.

Rosselli R, Romoli O, Vitulo N, et al. (12 co-authors). 2016. *Direct 16s rrna-seq from bacterial communities: a pcr-independent approach to simultaneously assess microbial diversity and functional activity potential of each taxon*. 6:32165.

Rosumek FB, Silveira FAO, de S Neves F, de U Barbosa NP, Diniz L, Oki Y, Pezzini F, Fernandes GW, Cornelissen T. 2009. Ants on plants: a meta-analysis of the role of ants as plant biotic defenses. *Oecologia*. 160:537–549.

Rubin BER, Moreau CS. 2016. Comparative genomics reveals convergent rates of evolution in ant-plant mutualisms. *Nature Communications*. 7:12679.

Rubinoff D, Cameron S, Will K. 2006. A genomic perspective on the shortcomings of mitochondrial dna for "barcoding" identification. *Journal of Heredity*. 97:581–594.

Rubinoff D, Holland BS. 2005. Between two extremes: Mitochondrial dna is neither the panacea nor the nemesis of phylogenetic and taxonomic inference. *Systematic Biology*. 54:952–961.

Rull V. 2011. Neotropical biodiversity: timing and potential drivers. *Trends in Ecology & Evolution*. 26:508–513.

Rull V, López-Sáez JA, Vegas-Vilarrúbia T. 2008. Contribution of non-pollen palynomorphs to the paleolimnological study of a high-altitude andean lake (laguna verde alta, venezuela). *Journal of Paleolimnology*. 40:399.

Saffo MB. 1992. Coming to terms with a field: words and concepts in symbiosis. *Symbiosis*. 14:17–31.

Safonova Y, Bankevich A, Pevzner PA. 2015. dipspades: Assembler for highly polymorphic diploid genomes. *Journal of computational biology : a journal of computational molecular cell biology*. 22:528–45.

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Molecular Biology and Evolution*. 31:1261–1271.

Sanchez A. 2015. Fidelity and promiscuity in an ant-plant mutualism: A case study of *Triplaris* and *Pseudomyrmex*. *PLOS ONE*. 10:1–19.

Sanders KL, Lee MSY. 2007. Evaluating molecular clock calibrations using bayesian analyses with soft and hard bounds. *Biology Letters*. 3:275—279.

Santos AMO, Jacobi CM, Silveira FAO. 2017. Frugivory and seed dispersal effectiveness in two *Miconia* (melastomataceae) species from ferruginous campo rupestre. *Seed Science Research*. 27:65–73.

Sayyari E, Mirarab S. 2016. Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC genomics*. 17:783.

Sayyari E, Whitfield JB, Mirarab S. 2017. Discovista: interpretable visualizations of gene tree discordance. *arXiv preprint arXiv:1709.09305*. .

Schatz MC, Delcher AL, Salzberg SL. 2010. Assembly of large genomes using second-generation sequencing. *Genome Research*. 20:1165–1173.

Schatz MC, Witkowski J, McCombie WR. 2012. Current challenges in de novo plant genome sequencing and assembly. *Genome biology*. 13:243.

Schindel DE, Miller SE. 2005. Dna barcoding a useful tool for taxonomists. *Nature*. 435:17–17.

Schmidt BC, Sperling FA. 2008. Widespread decoupling of mtdna variation and species integrity in *Grammia* tiger moths (lepidoptera: Noctuidae). *Systematic Entomology*. 33:613–634.

Schwartz MW, Hoeksema JD. 1998. Specialization and resource trade: biological markets as a model of mutualisms. *Ecology*. 79:1029–1038.

Seberg O. 2004. The future of systematics: assembling the tree of life. *Systematist*. pp. 2–8.

Segraves KA. 2010. Branching out with coevolutionary trees. *Evolution: Education and Outreach*. 3:62–70.

Seifert B. 2009. Cryptic species in ants (hymenoptera: Formicidae) revisited: we need a change in the alpha-taxonomic approach. *Myrmecological News*. 12:149–166.

Shaffer HB, Thomson RC, Weins J. 2007. Delimiting species in recent radiations. *Systematic Biology*. 56:896–906.

Shattuck SO. 1992. Generic revision of the ant subfamily dolichoderinae (hymenoptera: Formicidae). *Sociobiology*. 21:1–181.

Shattuck SO. 1995. Generic level relationships within the ant subfamily dolichoderinae (hymenoptera: Formicidae). *Systematic Entomology*. 20:217–228.

Shaw KL. 2002. Conflict between nuclear and mitochondrial dna phylogenies of a recent species radiation: What mtdna reveals and conceals about modes of speciation in hawaiian crickets. *Proceedings of the National Academy of Sciences*. 99:16122–16127.

Shephard GE, Muller RD, Liu L, Gurnis M. 2010. Miocene drainage reversal of the amazon river driven by plate-mantle interaction. *Nature Geosci*. 3:870–875.

Shik JZ, Kaspari M. 2009. Lifespan in male ants linked to mating syndrome. *Insectes Sociaux*. 56:131–134.

Silva GGZ, Dutilh BE, Matthews TD, Elkins K, Schmieder R, Dinsdale EA, Edwards RA. 2013. Combining *de novo* and reference-guided assembly with scaffold builder. *Source Code for Biology and Medicine*. 8:23.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. 31:3210–3212.

Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*. 99:13627–13632.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*. 15:121–132.

Smith AM, Fernández-Triana JL, Eveleigh E, et al. (11 co-authors). 2013. Dna barcoding and the taxonomy of microgastrinae wasps (hymenoptera, braconidae): impacts after 8 years and nearly 20,000 sequences. *Molecular Ecology Resources*. 13:168–176.

Smith MA, Bertrand C, Crosby K, et al. (28 co-authors). 2012. *Wolbachia* and dna barcoding insects: Patterns, potential, and problems. *PLOS ONE*. 7:e36514.

Smith MA, Fisher BL. 2009. Invasions, dna barcodes, and rapid biodiversity assessment using ants of mauritius. *Frontiers in Zoology*. 6:31.

Smith MA, Fisher BL, Hebert PDN. 2005. Dna barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: The ants of madagascar. *Philosophical Transactions: Biological Sciences*. 360:1825–1834.

Smith MA, Hallwachs W, Janzen DH. 2014. Diversity and phylogenetic community structure of ants along a costa rican elevational gradient. *Ecography*. 37:720–731.

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science*. 322:86–89.

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evolutionary Biology*. 15:150.

Solano PJ, Dejean A. 2004. Ant-fed plants: comparison between three geophytic myrmecophytes. *Biological Journal of the Linnean Society*. 83:433–439.

Solt ML, Wurdack JJ. 1980. Chromosome numbers in the melastomataceae. *Phytologia*. 47:199—220.

Solvestre R, Agosti D, Fernández F, et al. (11 co-authors). 2003. Introducción a las hormigas de la región Neotropical. Instituto de investigación de recursos biológicos Alexander von Humboldt.

Sorenson MD, Quinn TW. 1998. Numts: A challenge for avian systematics and population biology. *The Auk*. 115:214–221.

Sork VL, Fitz-Gibbon ST, Puiu D, Crepeau M, Gugger PF, Sherman R, Stevens K, Langley CH, Pellegrini M, Salzberg SL. 2016. First draft assembly and annotation of the genome of a california endemic oak *Quercus lobata* née (fagaceae). *G3: Genes, Genomes, Genetics*. 6:3485–3495.

Stahlhut JK, Fernández-Triana J, Adamowicz SJ, et al. (11 co-authors). 2013. Dna barcoding reveals diversity of hymenoptera and the dominance of parasitoids in a sub-arctic environment. *BMC ecology*. 13:2.

Stamatakis A. 2014. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–1313.

Stanton-Geddes J, Nguyen A, Chick L, Vincent J, Vangala M, Dunn RR, Ellison AM, Sanders NJ, Gotelli NJ, Cahan SH. 2016. Thermal reactionomes reveal divergent

responses to thermal extremes in warm and cool-climate ant species. *BMC Genomics*. 17:171.

Stork NE. 2018. How many species of insects and other terrestrial arthropods are there on earth? *Annual Review of Entomology*. 63:31–45. PMID: 28938083.

Sukumaran J, Holder MT. 2010. Dendropy: a python library for phylogenetic computing. *Bioinformatics*. 26:1569–1571.

Sukumaran J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. *Proceedings of the National Academy of Sciences*. .

Tänzler R, Sagata K, Surbakti S, Balke M, Riedel A. 2012. Dna barcoding for community ecology-how to tackle a hyperdiverse, mostly undescribed melanesian fauna. *PLoS One*. 7:e28832.

Thompson JN. 1994. The coevolutionary process. University of Chicago Press.

Thompson JN. 2005. The geographic mosaic of coevolution. University of Chicago Press.

Toews DPL, Brelsford A. 2012. The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*. 21:3907–3930.

Torres MF, Sanchez A. 2017. Neotropical ant-plant *Triplaris americana* attracts *Pseudomyrmex mordax* ant queens during seedling stages. *Insectes sociaux*. 64:255–261.

Treseder KK, Davidson DW, Ehleringer JR. 1995. Absorption of ant-provided carbon dioxide and nitrogen by a tropical epiphyte. *Nature*. 375:137.

Triana J. 1871. Les Melastomatacees, volume 28. Trans. Linn. Soc. London.

Troya A. 2012. Speciation of ants in the tropics of South America. Master's thesis, Universidad de Munich/2012.

Ueda S, Nagano Y, Kataoka Y, Komatsu T, Itioka T, Shimizu-kaya U, Inui Y, Itino T. 2015. Congruence of microsatellite and mitochondrial dna variation in acrobat ants (*Crematogaster* subgenus decacrema, formicidae: Myrmicinae) inhabiting *Macaranga* (euphorbiaceae) myrmecophytes. *PLOS ONE*. 10:1–15.

Ueda S, Quek SP, Itioka T, Inamori K, Sato Y, Murase K, Itino T. 2008. An ancient tripartite symbiosis of plants, ants and scale insects. *Proceedings of the Royal Society of London B: Biological Sciences*. 275:2319–2326.

Unamba CIN, Nag A, Sharma RK. 2015. Next generation sequencing technologies: The doorway to the unexplored genomics of non-model plants. *Frontiers in Plant Science*. 6:1074.

Van der Auwera GA, Carneiro MO, Hartl C, et al. (15 co-authors). 2002. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline, John Wiley & Sons, Inc.

Van der Hammen T. 1956. A palynological systematic nomenclature. *Bol. Geol., Bogota*. 4:63–101.

Van der Hammen T, Werner J, Van Dommelen H. 1973. Palynological record of the upheaval of the northern andes: a study of the pliocene and lower quaternary of the colombian eastern cordillera and the early evolution of its high-andean biota. *Review of Palaeobotany and Palynology*. 16:14784–4281122.

Vasconcelos HL. 1991. Mutualism between *Maieta guianensis* aubl., a myrmecophytic melastome, and one of its ant inhabitants: ant protection against insect herbivores. *Oecologia*. 87:295–298.

Vienne DMD, Giraud T, Shykoff JA. 2007. When can host shifts produce congruent host and parasite phylogenies? a simulation approach. *Journal of Evolutionary Biology*. 20:1428–1438.

Virgilio M, Backeljau T, Nevado B, De Meyer M. 2010. Comparative performances of dna barcoding across insect orders. *BMC Bioinformatics*. 11:206.

Virgilio M, Jordaens K, Breman FC, Backeljau T, De Meyer M. 2012. Identifying insects with incomplete dna barcode libraries, african fruit flies (diptera: Tephritidae) as a test case. *PLOS ONE*. 7:1–8.

Vizek AL, Brusich M, Arévalo JE, Mora-Pineda G. 2012. The shift in altitudinal range of *Azteca* (hymenoptera: Formicidae) ants inhabiting *Cecropia* (urticaceae) trees in monteverde, costa rica. *Brenesia*. .

Vogel M, Bänfer G, Moog U, Weising K. 2003. Development and characterization of chloroplast microsatellite markers in *Macaranga* (euphorbiaceae). *Genome*. 46:845–857. PMID: 14608402.

von Hagen KB, Kadereit JW. 2001. The phylogeny of *Gentianella* (gentianaceae) and its colonization of the southern hemisphere as revealed by nuclear and chloroplast dna sequence variation. *Organisms Diversity & Evolution*. 1:61–79.

Wade MJ. 2007. The co-evolutionary genetics of ecological communities. *Nature Reviews Genetics*. 8:185.

Walker JF, Brown JW, Smith SA. 2017. Site and gene-wise likelihoods unmask influential outliers in phylogenomic analyses. *bioRxiv*. .

Ward PS, Brady SG, Fisher BL, Schultz TR. 2010. Phylogeny and biogeography of dolichoderine ants: Effects of data partitioning and relict taxa on historical inference. *Systematic Biology*. .

Ward PS, Branstetter MG. 2017. The acacia ants revisited: convergent evolution and biogeographic context in an iconic ant/plant mutualism. *Proceedings of the Royal Society of London B: Biological Sciences*. 284.

Ward RD, Holmes BH. 2007. An analysis of nucleotide and amino acid variability in the barcode region of cytochrome c oxidase i (cox1) in fishes. *Molecular Ecology Notes*. 7:899–907.

Waugh J. 2007. Dna barcoding in animal species: progress, potential and pitfalls. *Bioessays*. 29:188–197.

Webber B, Joachim M, Curtis ASO, Woodrow IE. 2007. The diversity of ant-plant interactions in the rainforest understorey tree, *Ryparosa* (achariaceae): food bodies, domatia, prostomata, and hemipteran trophobionts. *Botanical Journal of the Linnean Society*. 154:353–371.

Weber NA. 1966. Fungus-growing ants. *Science*. 153:587.

Weeks AR, Turelli M, Harcombe WR, Reynolds KT, Hoffmann AA. 2007. From parasite to mutualist: Rapid evolution of wolbachia in natural populations of *Drosophila*. *PLOS Biology*. 5:e114.

Werren JH. 1997. Biology of *Wolbachia*. *Annual review of entomology*. 42:587–609.

Werren JH, Baldo L, Clark ME. 2008. *Wolbachia*: master manipulators of invertebrate biology. *Nat Rev Micro*. 6:741–751.

Werren JH, Zhang W, Guo LR. 1995. Evolution and phylogeny of *Wolbachia*: reproductive parasites of arthropods. *Proceedings of the Royal Society of London B: Biological Sciences*. 261:55–63.

Wheeler WM. 1912. A list of the type species of the genera and subgenera of formicidae. *Annals of the New York Academy of Sciences*. 21:157–175.

Wild AL. 2009. Evolution of the neotropical ant genus *Linepithema*. *Systematic Entomology*. 34:49—62.

Will KW, Mishler BD, Wheeler QD, Savolainen V. 2005. The perils of dna barcoding and the need for integrative taxonomy. *Systematic Biology*. 54:844–851.

Wilson EO. 1985. Ants of the dominican amber (hymenoptera: Formicidae). 3. the subfamily dolichoderinae. *Psyche*. 92:17–38.

Wilson EO, Hölldobler B. 2005. Eusociality: Origin and consequences. *Proceedings of the National Academy of Sciences of the United States of America*. 102:13367–13371.

Winterton C, Richardson JE, Hollingsworth M, Clark A, Zamora N, Pennington RT. 2014. Historical biogeography of the neotropical legume genus *Dussia*: the andes, the panama isthmus and the chocó. *Paleobotany and Biogeography: A Festschrift for Alan Graham in his 80th Year*. pp. 389–404.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution*. 26:1–7.

Yang Z. 1998. On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*. 47:125–133.

Yang Z. 2002. Likelihood and bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*. 162:1811–1823.

Yang Z. 2015. The bpp program for species tree estimation and species delimitation. *Current Zoology*. 61:854—865.

Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*. 107:9264–9269.

Yang Z, Rannala B. 2014. Unguided species delimitation using dna sequence data from multiple loci. *Molecular Biology and Evolution*. 31:3125–3135.

Yang Z, Rannala B. 2016. Species identification by bayesian fingerprinting: A powerful alternative to dna barcoding. *bioRxiv*. .

Yao H, Song J, Liu C, et al. (11 co-authors). 2010. Use of its2 region as the universal dna barcode for plants and animals. *PLOS ONE*. 5:e13102.

Yu DW, Davidson DW. 1997. Experimental studies of species-specificity in *Cecropia*-ant relationships. *Ecological Monographs*. 67:273–294.

Yu DW, Wilson HB, Frederickson ME, Palomin W, de la Colina R, Edwards DP, Balareso AA. 2004. Experimental demonstration of species coexistence enabled by dispersal limitation. *Journal of Animal Ecology*. 73:1102–1114.

Yu DW, Wilson HB, Pierce NE. 2001. An empirical model of species coexistence in a spatially structured environment. *Ecology*. 82:1761–1771.

Yu Y, Harris A, Blair C, He X. 2015. Rasp (reconstruct ancestral state in phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution*. 87:46–49.

Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, Landherr LL, Soltis DE, dePamphilis CW, Ma H. 2005. The evolution of the sepallata subfamily of mads-box genes. *Genetics*. 169:2209–2223.

Zambrano E, Vasquez E, Duval B, Latreille M, Coffinieres B. 1971. Síntesis paleogeográfica y petrolera del occidente de venezuela. *IVth Congr. Geol. Venez. Mem.* 1:483–552.

Zerbino DR, Birney E. 2008. Velvet: Algorithms for de novo short read assembly using de bruijn graphs. *Genome Research*. 18:821–829.

Zhang A, Muster C, Liang H, Zhu C, Crozier R, Wan P, Feng J, Ward RD. 2012. A fuzzy-set-theory-based approach to analyse species membership in dna barcoding. *Molecular Ecology*. 21:1848–1863.

Zhang J, Kapli P, Pavlidis P, Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*. 29:2869–2876.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. 2006. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Research*. 16:1099–1108.

Zhou Y, Duvaux L, Ren G, Zhang L, Savolainen O, Liu J. 2016. Importance of incomplete lineage sorting and introgression in the origin of shared genetic variation between two closely related pines with overlapping distributions. *Heredity*. 118:211–220.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The masurca genome assembler. *Bioinformatics*. 29:2669–2677.

Zimmermann T, Mirarab S, Warnow T. 2014. Bbca: Improving the scalability of* beast using random binning. *BMC genomics*. 15:S11.

Appendices

APPENDIX

A

APPENDIX INTRODUCTION

Literature review

This is a summary of the papers passing filters after the term search "'phylogeograph* AND (plant* AND ant*) OR myrmecoph*" in PubMed. For this search, the **Adjuntant R v.3.4.0** package was used (Crisan et al., 2018). The resulting database with 415 papers was first manually filtered by the title such that papers regarding obligate myrmecophytism will pass (ants inhabiting inside plants as opposed to ants visiting nectaries). Those papers (150) were subsequently filtered based on the Abstract such that it includes those with an evolutionary focus that was not restricted to ecological experiments only. Papers addressing third parties (fungi or scale insects) were included in case the phylogeography of the ants or plants could be extracted. This revision demonstrates that there is a small number of papers sampling both, ants and plants, but also that efforts are increasing with time.

From the papers passing the filters, I present in this order and separated by a semicolon: **Citation; Title; Plant species; Ant species; Collecting efforts; Geographic region; Main aim; Main conclusions; Geographic focus.**. DNA markers used in the studies are in brackets. Cladograms estimated during the studies for particular taxa only are represented by an asterisk. Taxa highlighted in bold represent the focus of the study.

Baker et al. (2017); Distinctive fungal communities in an obligate African ant-plant mutualism; *Vachellia* (*Acacia*) *drepanolobium*; *Tetraponera penzigi*, *Crematogaster nigriceps*, *Crematogaster mimosae*; **Fungal** samples collected from domatia and alate ants; Kenia; Description of fungal communities for *Vachellia*-ant associations; Fungal communities are specific to ant species, possibly due to ant behavior in

the domatia and by ants vectoring fungi when they disperse to establish new colonies; Distance-based ordinations to test differences in fungal community compositions across locations but geographic structure is not determinant.

Banfer et al. (2006); A chloroplast genealogy of myrmecophytic *Macaranga* species (Euphorbiaceae) in Southeast Asia reveals hybridization, vicariance and long-distance dispersals; *Macaranga* (atpB-rbcL, microsatellites); *Crematogaster*; Plants collected; Southeast Asia; Population genetics and phylogeography of ant and plant; Three myrmecophyte *Macaranga* form distinct clades in the network analysis. Structure of the chloroplast haplotypes is determined by geography rather than taxonomy. Myrmecophyte lineages originated at or had a continuous distribution in the Malay peninsula with subsequent colonization of Borneo, where geographic structure is also observed (northeast against the rest of Borneo due to old refugia on the rainforest); Population genetics and phylogeography of both, ant and plant.

Blatrix et al. (2013); Repeated evolution of fungal cultivar specificity in independently evolved ant-plant-fungus symbioses; *Leonardoxa africana africana*, *L. africana letouzeyi*, *Barteria fistulosa*; *Petalomyrmex phylax*, *Aphomomyrmex afer*, *Tetraponera aethiops*; In the study, they sequenced the **fungi** inside domatia from those plants (ITS); Fungal samples collected*; Cameroon, Nigeria, Lower Guinea, Congo; Description of fungal associations in three myrmecophyte systems; Each symbiosis was associated with a specific, dominant, primary fungal taxon, with one or two specific secondary taxa, all of the order Chaetothyriales. No geographic structure in fungi communities; No directly addressed.

Blattner et al. (2001); Molecular analysis of phylogenetic relationships among Myrmecophytic *Macaranga* species (Euphorbiaceae); *Macaranga* species

(ITS and RAPDs)*; *Crematogaster*, *Camponotus*; Plants collected on the field and at greenhouses; Southeast Asia; Phylogenetic relationships within *Macaranga* and the evolution of myrmecophytism; Ambiguous reconstruction of the origin of myrmecophytism. When ITS and RAPDs are combined, analyses reveal several independent gains and losses of myrmecophytism; No geographic structure analyses.

Chenuil and McKey (1996); Molecular phylogenetic study of a myrmecophyte symbiosis: did *Leonardoxa*/ ant associations diversify via cospeciation?; Four allelopatric lineages of ***Leonardoxa africana*** (ITS)*; ***Aphomomyrmex afer*, *Petalomyrmex phylax*** (COI and COII partial, 16S rDNA partial)*; Both collected and from museum collections, but is not clear if ant and plant collections were simultaneous; Cameroon; Testing for cospeciation between ants and plants; *A. after* and *P. phylax* colonized independently different lineages of *Leonardoxa*; No geographic structure analyses.

Chomicki and Renner (2015); Phylogenetics and molecular clocks reveal the repeated evolution of ant-plants after the late Miocene in Africa and the early Miocene in Australasia and the Neotropics; 681 myrmecophyte **plants** and their non-myrmecophyte sister taxa (18S rDNA, ITS, rbcL, matK, ndhF, atpB, trnL-trnF, atpB-rbcL)*; All phytophile **ants**; Sequences previously available; Global; Global patterns of evolution of myrmecophytism; Ant-plant symbioses evolved in the tropics. Overall c. 7 times more ant-plant species than plant-ant species. No ant-plant crown age is older than 19 Million Years (Myr) (Australasia), 15 Myr in Neotropics, and the oldest domatium-bearing species groups in Africa date to 6 Myr; Global patterns of evolution of myrmecophytism, but phylogeography is not directly addressed.

Chomicki and Renner (2016); Evolutionary relationships and biogeography of the ant-epiphytic genus *Squamellaria* (Rubiaceae: Psychotrieae) and their taxonomic implications. Hydnophytinae plants, ***Squamellaria*** species (trnL, trnL-trnF, ndhF, rps12-rpl20, trnS-trnG, rps16, 18S, ITS, ETS)*; *Philidris*, *Anonychomyrma*; Plants collected; Fiji, Vanuatu, the Solomons; Solving the phylogenetic placement of Hydnophytinae plants and 5 new myrmecophyte *Squamellaria* species; Some geographic structure within myrmecophyte *Squamellaria*, whose origin is dated 15-16 Myr. Most recent common ancestor of *Squamellaria* lived in Fiji and Vanuatu and subsequently colonized the Solomon Islands; Places the new species in a geographic context.

Davies et al. (2001); Evolution of myrmecophytism in western Malesian *Macaranga* (Euphorbiaceae); *Macaranga* species (ITS and morphological characters)*; *Crematogaster*, *Camponotus*; Plants from herbaria; Western Malesia; Evolution of myrmecophytism in *Macaranga*, how many times it appeared and possible restrictions to the evolution of myrmecophytism due to biogeographic factors; Myrmecophytism evolved multiple times in *Macaranga*, each time involving different sets of traits. Myrmecophyte *Macaranga* are restricted to areas with no seasonality; No geographic structure analyses.

Dev et al. (2010); Genetic and clonal diversity of the endemic ant-plant *Humboldtia brunonis* (Fabaceae) in the Western Ghats of India; *Humboldtia brunonis* (inter-simple sequence repeats - ISSRs)*; Not mentioned in the paper; Plants collected; South India; Genetic diversity and structure of clonal and plant populations. Geographic and genetic distances are correlated. Multiclonal populations exhibit low genetic diversity; Geographic structure between and within populations.

Gómez-Acevedo et al. (2010); Neotropical mutualism between *Acacia* and *Pseudomyrmex*: phylogeny and divergence times; *Acacia* (matK, psaB-rps14, trnL-trnF)*. ***Pseudomyrmex*** (LR, Wnt)*. Plants collected on the field and from herbaria. Ants collected; Sequences previously available; Mexico, including taxa from the Neotropics and Paleotropics; Elucidating the evolutionary history of the *Acacia-Pseudomyrmex* association. Crown *Acacia* date 5.44 Myr while their ants date 4.58 Myr. Their relationship originated in Mesoamerica between the late Miocene to the middle Pliocene, with eventual diversification of both groups in Mexico. Neotropical *Pseudomyrmex* is monophyletic; No directly addressed.

Heil et al. (2009); Divergent investment strategies of *Acacia* myrmecophytes and the coexistence of mutualists and exploiters; *Acacia cornigera*, *A. hindsii*, *A. collinsii*, *A. chiapensis* (trnK intron and trnL-trnF)*; ***Pseudomyrmex gracilis*, *P. ferrugineus*, *P. mixtecus*, *P. peperi*** (wg, abd-A,LW Rh, 28S, COI)*; Both plants and ants collected; Mexico; Effects of host reward production in the exploitation of the host by mutualist ants; High-reward hosts were better protected from herbivory and exploitation than hosts producing less rewards. No cheaters could be detected; No directly addressed though geographic structure on ant species is observed in the cladograms.

Leotard et al. (2008); Mutualism, hybrid unviability and speciation in a tropical ant-plant; *Leonardoxa africana africana* and ***L. africana gracilicaulis*** (six microsatellite markers); *Petalomyrmex phylax*; Plants collected throughout the hybridization zone; Cameroon; Hybridization between lineages of *Leonardoxa* and potential role of myrmecophytism in speciation; No introgression between subspecies but hybrids are F1 from both *Leonardoxa* subspecies, with intermediate traits but still

different from parental ones, especially regarding the number of nectaries and the domatia development. *P. phylax* was not found occupying hybrids despite being an obligate ant; No directly addressed.

Malé et al. (2016); Limited gene dispersal and spatial genetic structure as stabilizing factors in an ant-plant mutualism; *Hirtella physophora* (14 microsatellite markers)*; ***Allomerus decemarticulatus*** (10 microsatellite markers)*; Both collected; French Guyana; Population structure and stability of the mutualism; Significant fine-scale population genetic structure despite of closely related individuals being spatially close to each other. Moderate to strong genetic differentiation between sampling locations (mostly less than 50 km). The isolation by distance was highly significant for both. Short dispersal distances in both. All above is less marked for plants than ants; Assessment of the large-scale spatial genetic structure.

C. and L. (2009); Long-term persistence of a neotropical ant-plant population in the absence of obligate plant-ants; *Tococa guianensis*; ***Allomerus octoarticulatus***, ***Azteca***; Survey by transects of the plants and its ants, or ants identified from herbarium plant samples; Brazil; Geographic variation on ant defense in the *T. guianensis*-ant associations. Geographic differences on leaf traits and in the presence/absence of obligate and opportunist ants. Ants only absent in an altitudinally high location; Geographic variation of ants associated to *T. guianensis* was explored but *Azteca* was found in very low numbers and all of them in the same transect.

Nakashima et al. (2016); Morphological and phylogenetic investigations for several cryptic ant-plants found in *Callicarpa* (Lamiaceae) from Borneo; ***Callicarpa*** (ITS, matK, trnD-trnT)*; *Crematogaster*, *Technomyrmex*; Plants collected and sequences previously available; Indonesia; Plant morphology and ant inhabitation

to elucidate the evolution of myrmecophytism in *Callicarpa*; Two species with newly described domatia, confirmation of other *Callicarpa* species as myrmecophytes; No directly addressed.

Peccoud et al. (2013); Multi-locus phylogenies of the genus *Barteria* (Passifloraceae) portray complex patterns in the evolution of myrmecophytism; ***Barteria*** (ITS, 11 microsatellites)*; *Tetraponera*, *Crematogaster*; Plants collected and from herbaria; Cameroon, Gabon; Relationships within *Barteria* and the evolution of myrmecophyte traits on the genus; Myrmecophytism appears to be ancestral for myrmecophyte *Barteria* with subsequent loses of the mutualism. *Barteria* is a monophyletic genus although this is not demonstrated for its species. Geographic distance does not explain alone the genetic distance between species, suggesting additional barriers to gene flow must exist; Searched for correlations between geographic and genetic distances but correlation is not strong.

Pringle et al. (2012); Diversification and phylogeographic structure in widespread *Azteca* plant-ants from the northern Neotropics; *Cordia alliodora*; ***Azteca*** (ITS2, CO1, EF1aF1, LWRh, wg)*; Ants collected; Central America; Timing and geographic structure of the *Azteca-Cordia* mutualisms in Central America; *C. alliodora*-dwelling *Azteca* shared a common ancestor approximately 10–22 Million years Ago (Mya), before the proposed arrival of the host to Central America. Geographic structure within *A. pittieri* and no correlation between geographic and genetic distances, indicating that historical geologic or climatic conditions created barriers to gene flow; Geographic structure within *A. pittieri*, the most common obligate *Cordia* inhabitant.

Quek et al. (2004); Codiversification in an ant-plant mutualism: stem texture and the evolution of host use in *Crematogaster* (Formicidae: Myrmicinae) inhabitants of *Macaranga* (Euphorbiaceae); *Macaranga*; ***Crematogaster*** (COI)*. Both collected but the study uses a plant phylogeny already available; Southeast Asia; Testing for parallel cladogenesis (cospeciation) in *Macaranga* and their *Crematogaster* (Decacrema) ants; COI Bornean vs. out-of-Borneo geographic structure in ants. Confirmation of Myrmecophyte *Macaranga* species restricted to areas with no seasonality; Geographic structure in plants.

Razafimandimbison et al. (2005); Re-assessment of monophyly, evolution of myrmecophytism, and rapid radiation in *Neonauclea* s.s. (Rubiaceae); ***Neonauclea*** (ITS, ETS)*. *Cladomyrma*, *Crematogaster*; Plants collected or from herbaria; Southeast Asia; Malaysia; Phylogenetic relationships between *Neonauclea* and sister taxa and the evolution of myrmecophytism; Myrmecophytism evolved three times within the subfamily Naucleeae. *Neocauclea* and the sister *Myrmeconauclea* had non-myrmecophyte ancestors and the character was gained multiple independent times. Myrmecophytes are inhabited by different ant taxa in different geographic regions; Testing if myrmecophytism evolves once in *Neonauclea* as evidence shows that myrmecophyte species are geographically restricted.

Sanchez (2015); Fidelity and Promiscuity in an Ant-Plant Mutualism: A Case Study of *Triplaris* and *Pseudomyrmex*; ***Triplaris*** (psbA-trnH, rps16-trnK, nrITS, lfy2i)*. ***Pseudomyrmex*** (COI, LR)*; Both collected; Costa Rica, Colombia, Ecuador, Peru, Brazil; Understanding the evolution of ant-plant interactions in *Triplaris* and its ants; *P. mordax* is restricted to the west of the Eastern Cordillera whilst the rest of *Pseudomyrmex* species occur east to it. Most ants show broader host

usage except for *P. dendroicus* which is faithful to a single species of *Triplaris*. Host usage is not specific at the species level and preferences may result from geographical or ecological sorting. The specificity of *P. dendroicus* could be based on chemical recognition; Model the geographic distribution of *Triplaris* and *Pseudomyrmex*.

Ueda et al. (2008); An ancient tripartite symbiosis of plants, ants and scale insects; *Macaranga*; *Crematogaster* and ***Coccus*** coccids (COI for Coccus). Scale insects mostly collected from the ant colonies reported in Quek et al. (2007)*; Southeast Asia; Origin of tripartite mutualisms (ant, plants, scale insects); No specificity for the coccids on ants or plants. The minimum age of *Macaranga*-associated coccids is 7-8 Myr whilst for ants is 15-22 Myr; Indirectly showing lack of geographic structure.

Ueda et al. (2015); Congruence of microsatellite and mitochondrial DNA variation in acrobat ants (*Crematogaster* subgenus Decacrema, Formicidae: Myrmicinae) inhabiting *Macaranga* (Euphorbiaceae) myrmecophytes; *Macaranga*. ***Crematogaster*** (5 single sequence repeats, COI)*; Ants collected; Malaysia; Solving the incongruence between *Macaranga* morphological species and molecular clades found in previous studies. High host-plant specificity on ants. Cryptic genetic assemblages exhibiting high specificity toward plant species within a single *Macaranga* mtDNA clade; No directly addressed.

Vogel et al. (2003); Development and characterization of chloroplast microsatellite markers in *Macaranga* (Euphorbiaceae); 1-3 individuals from 10 ***Macaranga*** species (chloroplast microsatellites)*; *Crematogaster*, *Camponotus*; Plants collected; Southeast Asia; Development of markers to study the evolution of myrmecophytism in *Macaranga*; Haplotypes are determined more by geography than taxonomy. Geographic structure between haplotypes in West Malaysia and Borneo, and within

species in Borneo; Population genetics and phylogeography of myrmecophyte *versus* non-myrmecophyte *Macaranga*.

APPENDIX

B

APPENDIX CHAPTER 2

B.1 Species delimitation software

Box B.1 Commonly used software for species delimitation and specimen identification.

TaxonDNA: No need for a guide tree. Calculates pairwise distances for conspecific and congeneric sequences and determines a threshold value based on the distance frequencies. Requires previous knowledge about the groups (Meier et al., 2006).

jMOTU: No need for a guide tree. Performs an all-against-all Megablast of pre-clustered sequences, then filters the matches by comparing the Megablast score to a user-defined similarity threshold. Robust in the presence of indels. Prior knowledge of the groups is not required and works with unaligned sequences (Jones et al., 2011).

Automatic Barcode Gap Discovery (ABGD): No need for a guide tree. Sorts sequences by recursively calculating pairwise distances, finding the barcode gap and splitting the sequences in groups until no further splitting is possible. Prior knowledge of the groups is not required. Because it works with aligned sequences, it is less robust in the presence of indels (Puillandre et al., 2012).

Bayesian Phylogenetics and Phylogeography (BPP): No need for a guide tree or *a priori* assignment of samples to groups. Explores the whole space of possible trees and sample clustering in a Bayesian framework using nearest-neighbor interchange moves Requires priors on root height and ancestral effective population size. Delimits clusters under the multispecies coalescence model (Yang and Rannala, 2014).

General Mixed Yule Coalescent (GMYC): Requires an ultrametric tree and depends on how accurate the tree is. Not recommended for a large number of genes or a small number of species (O'meara, 2009). Identifies changes in branching time intervals and classifies them between interspecific (diversification) or intraspecific (coalescent) processes (Pons et al., 2006).

Poisson tree processes (PTP): Requires a tree (no need to be ultrametric). Uses the number of substitutions to identify changes between intra and interspecific branching processes. Assumes a higher number of substitutions between species than within species (Zhang et al., 2013).

Division of Individuals into Species using Sequences and Epsilon-Collapsed Trees (DISSECT): No need for guide tree or *a priori* assignment of samples to clusters. Explores the whole space of possible trees and sample clustering in a Bayesian framework. Priors on total number of clusters (Jones et al., 2014).

Table B.1: Coordinates of plants and ants collected. Non-myrmecophyte plants or empty plants are not included. Undetermined ants could not have been unambiguously identified and their sequences failed.

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
Cecropia	*Azteca*	MFT31	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT33	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT34	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT35	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT36	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT37	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT38	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT39	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT43	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT44	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT46	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT47	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT48	Choco	Arusi	5.569686	-77.496139	141
Cecropia	*Azteca*	MFT07	Meta	Villavicencio	4.05555	-73.84805	1539
Cecropia	*Azteca*	MFT04	Meta	Villavicencio	4.068333	-73.6638	439
Cecropia	*Azteca*	MFT17	Tolima	Espinal	4.01861	-74.8977	300
Cecropia	*Azteca*	MFT14	Tolima	Espinal	4.067777	-74.88194	302
Cecropia	*Azteca*	MFT15	Tolima	Espinal	4.08472	-74.86555	302
Cecropia	*Azteca*	MFT19	Tolima	Espinal	4.017	-74.89888	303
Cecropia	*Azteca*	MFT21	Tolima	Espinal	4.10111	-74.88194	310
Cecropia	*Azteca*	MFT10	Tolima	Espinal	4.13388	-74.915	329
Cecropia	*Azteca*	MFT11	Tolima	Espinal	4.13388	-74.915	329
Cecropia	*Azteca*	MFT12	Tolima	Espinal	4.1333	-74.91444	329
Cecropia	*Azteca*	MFT13	Tolima	Espinal	4.13472	-74.91444	329
Cecropia	*Azteca*	MFT25	Tolima	Guayabal	5.019166	-74.91527	336

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
Cecropia	*Azteca*	MFT09	Tolima	Espinal	4.1505	-74.93222	345
Cecropia	*Azteca*	MFT27	Tolima	Mariquita	5.13527	-74.86611	593
Clidemia hetero-phylla	*NN*	MFT462	Putumayo	Villagarzon	1.06629	-76.62440	494
Clidemia hetero-phylla	*Pheidole*	MFT461	Putumayo	Villagarzon	1.06629	-76.62440	494
Conostegia	*Pheidole*	MFT236	Choco	Arusi	5.57259	-77.49948	38
Conostegia	*Pheidole*	MFT216	Choco	Arusi	5.56923	-77.50240	41
Conostegia	*Pheidole*	MFT219	Choco	Arusi	5.57333	-77.50087	59
Conostegia	*Pheidole*	MFT222	Choco	Arusi	5.57329	-77.50035	72
Conostegia	*Pheidole*	MFT250	Choco	Arusi	5.57504	-77.49733	89
Conostegia	*Pheidole*	MFT251	Choco	Arusi	5.57504	-77.49733	89
Conostegia	*Pheidole*	MFT252	Choco	Arusi	5.57504	-77.49733	89
Conostegia	*Pheidole*	MFT199	Choco	Arusi	5.57134	-77.50090	91
Conostegia	*Pheidole*	MFT196	Choco	Arusi	5.57143	-77.50074	99
Conostegia	*Pheidole*	MFT200	Choco	Arusi	5.57143	-77.50074	99
Conostegia	*Pheidole*	MFT201	Choco	Arusi	5.57143	-77.50074	99
Conostegia	*Pheidole*	MFT202	Choco	Arusi	5.57143	-77.50074	99
Conostegia	*Pheidole*	MFT259	Choco	Arusi	5.57599	-77.50022	116
Conostegia	*Pheidole*	MFT261	Choco	Arusi	5.57599	-77.50022	116
Conostegia	*Pheidole*	MFT240	Choco	Arusi	5.57160	-77.49809	65
Conostegia	*Pheidole*	MFT243	Choco	Arusi	5.57160	-77.49809	65
Conostegia	*Pheidole*	MFT203	Choco	Arusi	5.57137	-77.50076	
Conostegia	*Pheidole*	MFT322	Valle del Cauca	Buenaventura	3.95744	-77.01948	30
Conostegia	*Pheidole*	MFT323	Valle del Cauca	Buenaventura	3.95744	-77.01948	30
Conostegia	*Pheidole*	MFT324	Valle del Cauca	Buenaventura	3.95744	-77.01948	30
Conostegia	*Pheidole*	MFT325	Valle del Cauca	Buenaventura	3.95744	-77.01948	30

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
Conostegia	*Pheidole*	MFT326	Valle del Cauca	Buenaventura	3.95744	-77.01948	30
Conostegia	*Azteca*	MFT372	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
Conostegia	*NN*	MFT371	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
Conostegia	*NN*	MFT373	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
Conostegia	*NN*	MFT374	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
Conostegia setosa	*Pheidole*	MFT174	Choco	Arusi	5.57330	-77.50177	23
Conostegia setosa	*Pheidole*	MFT182	Choco	Arusi	5.57449	-77.50007	82
Conostegia setosa	*Pheidole*	MFT176	Choco	Arusi	5.57393	-77.50116	40
Conostegia setosa	*Pheidole*	MFT178	Choco	Arusi	5.57450	-77.50023	67
Henriettella cuneata	*Pheidole*	MFT210	Choco	Arusi	5.57136	-77.50111	53
Henriettella cuneata	*Pheidole*	MFT173	Choco	Arusi	5.574062	-77.501185	53
Henriettella cuneata	*Pheidole*	MFT208	Choco	Arusi	5.57151	-77.50118	53
Maieta	*NN*	MFT417	Putumayo	Villagarzon	1.05531	-76.62315	562
Maieta guianensis	*Allomerus*	MFT152	Amazonas	Leticia	-4.120167	-69.956450	93
Maieta guianensis	*Allomerus*	MFT142	Amazonas	Leticia	-4.120267	-69.955150	94
Maieta guianensis	*Allomerus*	MFT143	Amazonas	Leticia	-4.120267	-69.955150	94
Maieta guianensis	*Allomerus*	MFT144	Amazonas	Leticia	-4.120267	-69.955150	94
Maieta guianensis	*Allomerus*	MFT145	Amazonas	Leticia	-4.120267	-69.955150	94
Maieta guianensis	*Pheidole*	MFT464	Putumayo	Villagarzon	1.06509	-76.62222	529
Maieta guianensis	*Pheidole*	MFT465	Putumayo	Villagarzon	1.06509	-76.62222	529
Maieta poeppigii	*Azteca*	MFT158	Amazonas	Leticia	-4.096450	-69.936188	88
Maieta poeppigii	*Azteca*	MFT159	Amazonas	Leticia	-4.096450	-69.936188	88
Maieta poeppigii	*Azteca*	MFT161	Amazonas	Leticia	-4.096450	-69.936188	88
Maieta poeppigii	*Azteca*	MFT162	Amazonas	Leticia	-4.096450	-69.936188	88

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
Maieta poeppigii	*NN*	MFT160	Amazonas	Leticia	-4.096450	-69.936188	88
Miconia	*Pheidole*	MFT395	Antioquia	San Carlos	6.19437	-75.00548	1084
Miconia	*Pheidole*	MFT329	Valle del Cauca	Buenaventura	3.95731	-77.01983	27
Miconia	*Wasmania*	MFT370	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
Ossaea bullifera	*Pheidole*	MFT149	Amazonas	Leticia	-4.120033	-69.956267	92
Ossaea bullifera	*Crematogaster*	MFT156	Amazonas	Leticia	-4.120800	-69.953983	119
Ossaea bullifera	*Pheidole*	MFT155	Amazonas	Leticia	-4.120800	-69.953983	119
Tococa	*Azteca*	MFT397	Antioquia	San Carlos	6.19439	-75.00578	1111
Tococa	*Azteca*	MFT147	Amazonas	Leticia	-4.120033	-69.956267	92
Tococa	*Azteca*	MFT148	Amazonas	Leticia	-4.120033	-69.956267	92
Tococa	*Pheidole*	MFT205	Choco	Arusi	5.57137	-77.50076	64
Tococa	*Pheidole*	MFT188	Choco	Arusi	5.57469	-77.49978	77
Tococa	*Azteca*	MFT410	Putumayo	Villagarzon	1.06082	-76.62243	515
Tococa	*NN*	MFT460	Putumayo	Villagarzon	1.06447	-76.62798	579
Tococa	*Azteca*	MFT341	Valle del Cauca	Buenaventura	3.97011	-77.00326	66
T. bullifera	*Allomerus*	MFT412	Putumayo	Villagarzon	1.05934	-76.62243	479
T. bullifera	*Azteca*	MFT413	Putumayo	Villagarzon	1.05934	-76.62243	479
T. bullifera	*NN*	MFT414	Putumayo	Villagarzon	1.05934	-76.62243	479
T. bullifera	*Azteca*	MFT415	Putumayo	Villagarzon	1.05934	-76.62243	479
T. bullifera	*Azteca*	MFT416	Putumayo	Villagarzon	1.05934	-76.62243	479
T. bullifera	*Azteca*	MFT402	Putumayo	Villagarzon	1.06146	-76.62263	503
T. bullifera	*Azteca*	MFT403	Putumayo	Villagarzon	1.06146	-76.62263	503
T. bullifera	*Azteca*	MFT406	Putumayo	Villagarzon	1.06146	-76.62263	503
T. bullifera	*Azteca*	MFT404	Putumayo	Villagarzon	1.06146	-76.62263	503
T. bullifera	*NN*	MFT405	Putumayo	Villagarzon	1.06146	-76.62263	503
T. bullifera	*Azteca*	MFT407	Putumayo	Villagarzon	1.06082	-76.62243	515
T. bullifera	*Azteca*	MFT408	Putumayo	Villagarzon	1.06082	-76.62243	515

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. bullifera	*Azteca*	MFT409	Putumayo	Villagarzon	1.06082	-76.62243	515
T. bullifera	*Allomerus*	MFT411	Putumayo	Villagarzon	1.06082	-76.62243	515
T. bullifera	*Allomerus*	MFT441	Putumayo	Villagarzon	1.05942	-76.62523	550
T. bullifera	*Azteca*	MFT440	Putumayo	Villagarzon	1.05942	-76.62523	550
T. bullifera	*Azteca*	MFT444	Putumayo	Villagarzon	1.05942	-76.62523	550
T. bullifera	*Azteca*	MFT458	Putumayo	Villagarzon	1.05979	-76.62585	550
T. bullifera	*NN*	MFT442	Putumayo	Villagarzon	1.05942	-76.62523	550
T. bullifera	*Pheidole*	MFT443	Putumayo	Villagarzon	1.05942	-76.62523	550
T. bullifera	*Pheidole*	MFT459	Putumayo	Villagarzon	1.05979	-76.62585	550
T. bullifera	*Azteca*	MFT472	Putumayo	Villagarzon	1.06665	-76.62077	579
T. bullifera	*Pheidole*	MFT471	Putumayo	Villagarzon	1.06665	-76.62077	579
T. bullifera	*Azteca*	MFT474	Putumayo	Villagarzon	1.06686	-76.62064	582
T. bullifera	*Azteca*	MFT475	Putumayo	Villagarzon	1.06686	-76.62064	582
T. bullifera	*NN*	MFT476	Putumayo	Villagarzon	1.06686	-76.62064	582
T. bullifera	*NN*	MFT487	Putumayo	Villagarzon	1.06694	-76.62038	596
T. caquetana	*Myrmelachista*	MFT448	Putumayo	Villagarzon	1.06005	-76.62702	540
T. caquetana	*Myrmelachista*	MFT434	Putumayo	Villagarzon	1.05577	-76.62359	588
T. caquetana	*Myrmelachista*	MFT435	Putumayo	Villagarzon	1.05577	-76.62359	588
T. caquetana	*Myrmelachista*	MFT436	Putumayo	Villagarzon	1.05577	-76.62359	588
T. caquetana	*Myrmelachista*	MFT433	Putumayo	Villagarzon	1.05577	-76.62359	588
T. caquetana	*NN*	MFT437	Putumayo	Villagarzon	1.05577	-76.62359	588
T. caquetana	*Crematogaster*	MFT439	Putumayo	Villagarzon	1.05942	-76.62523	548
T. cordata	*Azteca*	MFT169	Amazonas	Puerto Nariño	-3.783100	-70.363483	70
T. cordata	*Azteca*	MFT170	Amazonas	Puerto Nariño	-3.783100	-70.363483	70
T. cordata	*Azteca*	MFT172	Amazonas	Puerto Nariño	-3.783100	-70.363483	70
T. cordata	*Camponotus*	MFT168	Amazonas	Puerto Nariño	-3.783100	-70.363483	70
T. cordata	*Nylanderia*	MFT171	Amazonas	Puerto Nariño	-3.783100	-70.363483	70

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. cordata	*Azteca*	MFT166	Amazonas	Leticia	-4.096450	-69.936188	88
T. cordata	*Azteca*	MFT164	Amazonas	Leticia	-4.096450	-69.936188	88
T. cordata	*Pheidole*	MFT163	Amazonas	Leticia	-4.096450	-69.936188	88
T. cordata	*Pheidole*	MFT165	Amazonas	Leticia	-4.096450	-69.936188	88
T. cordata	*Pheidole*	MFT167	Amazonas	Leticia	-4.096450	-69.936188	88
T. guianensis	*Azteca*	MFT150	Amazonas	Leticia	-4.120167	-69.956450	93
T. guianensis	*Azteca*	MFT151	Amazonas	Leticia	-4.120167	-69.956450	93
T. guianensis	*Azteca*	MFT153	Amazonas	Leticia	-4.121283	-69.956217	98
T. guianensis	*NN*	MFT154	Amazonas	Leticia	-4.121283	-69.956217	98
T. guianensis	*Azteca*	MFT146	Amazonas	Leticia	-4.120233	-69.955417	102
T. guianensis	*Crematogaster*	MFT376	Antioquia	San Luis	6.04841	-74.99118	938
T. guianensis	*Azteca*	MFT375	Antioquia	San Luis	6.04826	-74.99151	948
T. guianensis	*Pheidole*	MFT377	Antioquia	San Luis	6.04825	-74.99128	975
T. guianensis	*Azteca*	MFT385	Antioquia	San Luis	6.04834	-74.99032	985
T. guianensis	*Azteca*	MFT386	Antioquia	San Luis	6.04834	-74.99032	985
T. guianensis	*Azteca*	MFT388	Antioquia	San Luis	6.04813	-74.99098	988
T. guianensis	*Azteca*	MFT384	Antioquia	San Luis	6.04817	-74.99128	988
T. guianensis	*Azteca*	MFT378	Antioquia	San Luis	6.04817	-74.99128	988
T. guianensis	*Azteca*	MFT383	Antioquia	San Luis	6.04817	-74.99128	988
T. guianensis	*Wasmania*	MFT387	Antioquia	San Luis	6.04813	-74.99098	988
T. guianensis	*Azteca*	MFT390	Antioquia	San Carlos	6.19437	-75.00548	1084
T. guianensis	*Azteca*	MFT391	Antioquia	San Carlos	6.19437	-75.00548	1084
T. guianensis	*Azteca*	MFT393	Antioquia	San Carlos	6.19437	-75.00548	1084
T. guianensis	*Crematogaster*	MFT394	Antioquia	San Carlos	6.19437	-75.00548	1084
T. guianensis	*Pheidole*	MFT392	Antioquia	San Carlos	6.19437	-75.00548	1084
T. guianensis	*Wasmania*	MFT389	Antioquia	San Luis	6.19437	-75.00548	1084
T. guianensis	*Azteca*	MFT396	Antioquia	San Carlos	6.19439	-75.00578	1111

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT398	Antioquia	San Carlos	6.19439	-75.00578	1111
T. guianensis	*Azteca*	MFT574	Antioquia	Amalfi	6.95020	-74.90156	1115
T. guianensis	*Azteca*	MFT568	Antioquia	Amalfi	6.95101	-74.90107	1122
T. guianensis	*Azteca*	MFT572	Antioquia	Amalfi	6.95072	-74.90108	1122
T. guianensis	*Azteca*	MFT573	Antioquia	Amalfi	6.95072	-74.90108	1122
T. guianensis	*Azteca*	MFT569	Antioquia	Amalfi	6.95081	-74.90100	1126
T. guianensis	*Azteca*	MFT570	Antioquia	Amalfi	6.95081	-74.90100	1126
T. guianensis	*Azteca*	MFT571	Antioquia	Amalfi	6.95081	-74.90100	1126
T. guianensis	*Azteca*	MFT575	Antioquia	Amalfi	6.95841	-74.89703	1126
T. guianensis	*Azteca*	MFT576	Antioquia	Amalfi	6.95841	-74.89703	1126
T. guianensis	*Azteca*	MFT577	Antioquia	Amalfi	6.95841	-74.89703	1126
T. guianensis	*Azteca*	MFT578	Antioquia	Amalfi	6.95841	-74.89703	1126
T. guianensis	*Azteca*	MFT399	Antioquia	San Carlos	6.19391	-75.00517	1127
T. guianensis	*Azteca*	MFT401	Antioquia	San Carlos	6.19391	-75.00517	1127
T. guianensis	*Azteca*	MFT400	Antioquia	San Carlos	6.19391	-75.00517	1127
T. guianensis	*Azteca*	MFT557	Antioquia	Amalfi	6.94781	-74.89741	1128
T. guianensis	*Azteca*	MFT558	Antioquia	Amalfi	6.94779	-74.89740	1128
T. guianensis	*Azteca*	MFT582	Antioquia	Amalfi	6.95851	-74.89740	1128
T. guianensis	*Azteca*	MFT553	Antioquia	Amalfi	6.94823	-74.89793	1129
T. guianensis	*Azteca*	MFT554	Antioquia	Amalfi	6.94823	-74.89793	1129
T. guianensis	*Azteca*	MFT559	Antioquia	Amalfi	6.94753	-74.89717	1130
T. guianensis	*Azteca*	MFT564	Antioquia	Amalfi	6.94596	-74.89629	1134
T. guianensis	*Azteca*	MFT583	Antioquia	Amalfi	6.95938	-74.89647	1134
T. guianensis	*Azteca*	MFT584	Antioquia	Amalfi	6.95938	-74.89647	1134
T. guianensis	*Azteca*	MFT585	Antioquia	Amalfi	6.95938	-74.89647	1134
T. guianensis	*Azteca*	MFT591	Antioquia	Amalfi	6.95946	-74.89654	1134
T. guianensis	*Azteca*	MFT579	Antioquia	Amalfi	6.95853	-74.89793	1137

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT580	Antioquia	Amalfi	6.95853	-74.89793	1137
T. guianensis	*Azteca*	MFT581	Antioquia	Amalfi	6.95853	-74.89793	1137
T. guianensis	*Azteca*	MFT555	Antioquia	Amalfi	6.94804	-74.89778	1138
T. guianensis	*Azteca*	MFT556	Antioquia	Amalfi	6.94804	-74.89778	1138
T. guianensis	*Azteca*	MFT560	Antioquia	Amalfi	6.94704	-74.89685	1138
T. guianensis	*Azteca*	MFT561	Antioquia	Amalfi	6.94704	-74.89685	1138
T. guianensis	*Azteca*	MFT562	Antioquia	Amalfi	6.94704	-74.89685	1138
T. guianensis	*Azteca*	MFT563	Antioquia	Amalfi	6.94701	-74.89688	1141
T. guianensis	*Azteca*	MFT586	Antioquia	Amalfi	6.95934	-74.89647	1142
T. guianensis	*Azteca*	MFT587	Antioquia	Amalfi	6.95934	-74.89647	1142
T. guianensis	*Azteca*	MFT588	Antioquia	Amalfi	6.95934	-74.89647	1142
T. guianensis	*Azteca*	MFT589	Antioquia	Amalfi	6.95934	-74.89647	1142
T. guianensis	*Azteca*	MFT590	Antioquia	Amalfi	6.95934	-74.89647	1142
T. guianensis	*Azteca*	MFT565	Antioquia	Amalfi	6.94566	-74.89586	1156
T. guianensis	*Azteca*	MFT566	Antioquia	Amalfi	6.94529	-74.89529	1178
T. guianensis	*Azteca*	MFT567	Antioquia	Amalfi	6.94529	-74.89529	1178
T. guianensis	*Azteca*	MFT379	Antioquia	San Luis	6.04817	-74.99128	988
T. guianensis	*Azteca*	MFT381	Antioquia	San Luis	6.04817	-74.99128	988
T. guianensis	*Azteca*	MFT382	Antioquia	San Luis	6.04817	-74.99128	988
T. guianensis	*Pheidole*	MFT521	Casanare	Villanueva	4.61264	-72.92070	260
T. guianensis	*Pheidole*	MFT525	Casanare	Villanueva	4.61272	-72.92062	264
T. guianensis	*Azteca*	MFT526	Casanare	Villanueva	4.61239	-72.92076	265
T. guianensis	*Azteca*	MFT527	Casanare	Villanueva	4.61157	-72.92080	266
T. guianensis	*Pheidole*	MFT522	Casanare	Villanueva	4.61263	-72.92070	276
T. guianensis	*Camponotus*	MFT523	Casanare	Villanueva	4.61266	-72.92069	277
T. guianensis	*Azteca*	MFT524	Casanare	Villanueva	4.61272	-72.92067	278
T. guianensis	*Azteca*	MFT520	Casanare	Villanueva	4.61250	-72.92080	287

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT538	Casanare	Tauramena	5.00644	-72.75368	447
T. guianensis	*Azteca*	MFT539	Casanare	Tauramena	5.00660	-72.75414	451
T. guianensis	*Azteca*	MFT540	Casanare	Tauramena	5.00660	-72.75414	451
T. guianensis	*Azteca*	MFT541	Casanare	Tauramena	5.00660	-72.75414	451
T. guianensis	*Azteca*	MFT542	Casanare	Tauramena	5.00660	-72.75414	451
T. guianensis	*Azteca*	MFT543	Casanare	Tauramena	5.00660	-72.75414	451
T. guianensis	*Azteca*	MFT529	Casanare	Tauramena	5.00634	-72.75349	453
T. guianensis	*Azteca*	MFT536	Casanare	Tauramena	5.00634	-72.75349	453
T. guianensis	*Azteca*	MFT537	Casanare	Tauramena	5.00634	-72.75349	453
T. guianensis	*Azteca*	MFT528	Casanare	Tauramena	5.02006	-72.76419	497
T. guianensis	*Azteca*	MFT531	Casanare	Tauramena	5.01998	-72.76491	497
T. guianensis	*Azteca*	MFT532	Casanare	Tauramena	5.02003	-72.76415	498
T. guianensis	*Azteca*	MFT533	Casanare	Tauramena	5.02003	-72.76415	498
T. guianensis	*Azteca*	MFT534	Casanare	Tauramena	5.02003	-72.76415	498
T. guianensis	*Azteca*	MFT535	Casanare	Tauramena	5.02003	-72.76415	498
T. guianensis	*Pheidole*	MFT530	Casanare	Tauramena	5.01999	-72.76447	498
T. guianensis	*Pheidole*	MFT193	Choco	Arusi	5.57169	-77.50067	38
T. guianensis	*Pheidole*	MFT237	Choco	Arusi	5.57259	-77.49948	38
T. guianensis	*Pheidole*	MFT185	Choco	Arusi	5.57435	-77.50003	59
T. guianensis	*Pheidole*	MFT220	Choco	Arusi	5.57335	-77.50061	66
T. guianensis	*Pheidole*	MFT223	Choco	Arusi	5.57331	-77.50000	66
T. guianensis	*Pheidole*	MFT179	Choco	Arusi	5.57450	-77.50023	67
T. guianensis	*Pheidole*	MFT180	Choco	Arusi	5.57450	-77.50023	67
T. guianensis	*Pheidole*	MFT226	Choco	Arusi	5.57283	-77.49967	71
T. guianensis	*Pheidole*	MFT186	Choco	Arusi	5.57441	-77.49999	72
T. guianensis	*Pheidole*	MFT225	Choco	Arusi	5.57316	-77.49990	73
T. guianensis	*Pheidole*	MFT183	Choco	Arusi	5.57438	-77.50011	74

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Pheidole*	MFT187	Choco	Arusi	5.57455	-77.50023	76
T. guianensis	*Pheidole*	MFT177	Choco	Arusi	5.57433	-77.50130	85
T. guianensis	*Pheidole*	MFT248	Choco	Arusi	5.57504	-77.49733	89
T. guianensis	*Pheidole*	MFT249	Choco	Arusi	5.57504	-77.49733	89
T. guianensis	*Tapinoma*	MFT247	Choco	Arusi	5.57504	-77.49733	89
T. guianensis	*Pheidole*	MFT197	Choco	Arusi	5.57143	-77.50074	99
T. guianensis	*Tapinoma*	MFT198	Choco	Arusi	5.57143	-77.50074	99
T. guianensis	*Pheidole*	MFT255	Choco	Arusi	5.57532	-77.49722	107
T. guianensis	*Pheidole*	MFT258	Choco	Arusi	5.57532	-77.49722	107
T. guianensis	*Pheidole*	MFT194	Choco	Arusi	5.57147	-77.50039	40
T. guianensis	*Pheidole*	MFT206	Choco	Arusi	5.57151	-77.50118	53
T. guianensis	*Pheidole*	MFT209	Choco	Arusi	5.57136	-77.50111	53
T. guianensis	*Pheidole*	MFT211	Choco	Arusi	5.57136	-77.50111	53
T. guianensis	*Pheidole*	MFT181	Choco	Arusi	5.57435	-77.50003	59
T. guianensis	*Pheidole*	MFT217	Choco	Arusi	5.57165	-77.50201	60
T. guianensis	*Tapinoma*	MFT218	Choco	Arusi	5.57165	-77.50201	60
T. guianensis	*Pheidole*	MFT215	Choco	Arusi	5.56961	-77.50166	62
T. guianensis	*Pheidole*	MFT238	Choco	Arusi	5.57160	-77.49809	65
T. guianensis	*Pheidole*	MFT239	Choco	Arusi	5.57160	-77.49809	65
T. guianensis	*Pheidole*	MFT244	Choco	Arusi	5.57160	-77.49809	65
T. guianensis	*Pheidole*	MFT245	Choco	Arusi	5.57160	-77.49809	65
T. guianensis	*Pheidole*	MFT246	Choco	Arusi	5.57160	-77.49809	65
T. guianensis	*Pheidole*	MFT230	Choco	Arusi	5.57283	-77.49967	71
T. guianensis	*Pheidole*	MFT227	Choco	Arusi	5.57283	-77.49967	71
T. guianensis	*Pheidole*	MFT228	Choco	Arusi	5.57283	-77.49967	71
T. guianensis	*Pheidole*	MFT229	Choco	Arusi	5.57283	-77.49967	71
T. guianensis	*Pheidole*	MFT231	Choco	Arusi	5.57283	-77.49967	71

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Pheidole*	MFT232	Choco	Arusi	5.57283	-77.49967	71
T. guianensis	*Pheidole*	MFT234	Choco	Arusi	5.57283	-77.49967	71
T. guianensis	*Pheidole*	MFT221	Choco	Arusi	5.57329	-77.50035	72
T. guianensis	*Pheidole*	MFT195	Choco	Arusi	5.57143	-77.50074	99
T. guianensis	*Azteca*	MFT272	Meta	S.J. de Arama	3.34677	-73.94125	471
T. guianensis	*Azteca*	MFT273	Meta	S.J. de Arama	3.34680	-73.94115	485
T. guianensis	*Azteca*	MFT274	Meta	S.J. de Arama	3.34680	-73.94115	485
T. guianensis	*Azteca*	MFT275	Meta	S.J. de Arama	3.34680	-73.94115	485
T. guianensis	*Azteca*	MFT276	Meta	S.J. de Arama	3.34680	-73.94115	485
T. guianensis	*Azteca*	MFT271	Meta	S.J. de Arama	3.35024	-73.93856	503
T. guianensis	*Azteca*	MFT277	Meta	S.J. de Arama	3.34654	-73.94096	505
T. guianensis	*Azteca*	MFT268	Meta	S.J. de Arama	3.35027	-73.93858	506
T. guianensis	*Azteca*	MFT270	Meta	S.J. de Arama	3.35032	-73.93868	510
T. guianensis	*Azteca*	MFT266	Meta	S.J. de Arama	3.35989	-73.95729	537
T. guianensis	*Azteca*	MFT267	Meta	S.J. de Arama	3.35628	-73.95497	667
T. guianensis	*Azteca*	MFT296	Meta	S.J. de Arama	3.273785	-73.893999	361
T. guianensis	*Azteca*	MFT299	Meta	S.J. de Arama	3.273785	-73.893999	361
T. guianensis	*Azteca*	MFT297	Meta	S.J. de Arama	3.273785	-73.893999	361
T. guianensis	*Azteca*	MFT301	Meta	S.J. de Arama	3.273785	-73.893999	361
T. guianensis	*Azteca*	MFT298	Meta	S.J. de Arama	3.273785	-73.893999	361
T. guianensis	*Azteca*	MFT300	Meta	S.J. de Arama	3.273785	-73.893999	361
T. guianensis	*Solenopsis*	MFT295	Meta	S.J. de Arama	3.273955	-73.894612	361
T. guianensis	*Azteca*	MFT290	Meta	S.J. de Arama	3.273785	-73.893999	362
T. guianensis	*Azteca*	MFT291	Meta	S.J. de Arama	3.273785	-73.893999	362
T. guianensis	*Solenopsis*	MFT292	Meta	S.J. de Arama	3.273785	-73.893999	362
T. guianensis	*Azteca*	MFT284	Meta	S.J. de Arama	3.252329	-73.947661	398
T. guianensis	*Azteca*	MFT287	Meta	S.J. de Arama	3.252329	-73.947661	398

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT283	Meta	S.J. de Arama	3.252329	-73.947661	398
T. guianensis	*Azteca*	MFT286	Meta	S.J. de Arama	3.252329	-73.947661	398
T. guianensis	*Azteca*	MFT285	Meta	S.J. de Arama	3.252329	-73.947661	398
T. guianensis	*Azteca*	MFT288	Meta	S.J. de Arama	3.274046	-73.920469	421
T. guianensis	*Myrmelachista*	MFT289	Meta	S.J. de Arama	3.274046	-73.920469	421
T. guianensis	*Azteca*	MFT511	Meta	Villavicencio	4.10028	-73.66082	465
T. guianensis	*Azteca*	MFT509	Meta	Villavicencio	4.10033	-73.66080	469
T. guianensis	*Azteca*	MFT546	Meta	Villavicencio	4.17798	-73.63616	469
T. guianensis	*Azteca*	MFT547	Meta	Villavicencio	4.17798	-73.63616	469
T. guianensis	*Azteca*	MFT549	Meta	Villavicencio	4.17798	-73.63616	469
T. guianensis	*Azteca*	MFT550	Meta	Villavicencio	4.17798	-73.63616	469
T. guianensis	*Azteca*	MFT551	Meta	Villavicencio	4.17798	-73.63616	469
T. guianensis	*Azteca*	MFT552	Meta	Villavicencio	4.17798	-73.63616	469
T. guianensis	*Azteca*	MFT548	Meta	Villavicencio	4.17798	-73.63616	469
T. guianensis	*Azteca*	MFT545	Meta	Villavicencio	4.17599	-73.63703	471
T. guianensis	*Pheidole*	MFT544	Meta	Villavicencio	4.17577	-73.63702	471
T. guianensis	*Azteca*	MFT508	Meta	Villavicencio	4.10047	-73.66081	473
T. guianensis	*Azteca*	MFT510	Meta	Villavicencio	4.10021	-73.66077	476
T. guianensis	*Azteca*	MFT512	Meta	Villavicencio	4.10045	-73.66075	477
T. guianensis	*Azteca*	MFT263	Meta	S.J. de Arama	3.35990	-73.05728	547
T. guianensis	*Azteca*	MFT264	Meta	S.J. de Arama	3.35990	-73.05728	547
T. guianensis	*Azteca*	MFT265	Meta	S.J. de Arama	3.35990	-73.05728	547
T. guianensis	*Azteca*	MFT262	Meta	S.J. de Arama	3.35990	-73.05728	547
T. guianensis	*Azteca*	MFT514	Meta	Acacías	3.96085	-73.78100	552
T. guianensis	*Azteca*	MFT515	Meta	Acacías	3.96085	-73.78100	552
T. guianensis	*Azteca*	MFT516	Meta	Acacías	3.96085	-73.78100	552
T. guianensis	*Azteca*	MFT517	Meta	Acacías	3.96085	-73.78100	552

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT518	Meta	Acacías	3.96085	-73.78100	552
T. guianensis	*Azteca*	MFT519	Meta	Acacías	3.96085	-73.78100	552
T. guianensis	*Azteca*	MFT513	Meta	Acacías	4.03152	-73.77449	555
T. guianensis	*Azteca*	MFT278	Meta	S.J. de Arama	3.255973	-73.959157	642
T. guianensis	*Azteca*	MFT280	Meta	S.J. de Arama	3.255973	-73.959157	642
T. guianensis	*Azteca*	MFT281	Meta	S.J. de Arama	3.255973	-73.959157	642
T. guianensis	*Azteca*	MFT282	Meta	S.J. de Arama	3.255973	-73.959157	642
T. guianensis	*Azteca*	MFT279	Meta	S.J. de Arama	3.255973	-73.959157	642
T. guianensis	*Azteca*	MFT455	Putumayo	Villagarzon	1.06642	-76.62433	482
T. guianensis	*NN*	MFT453	Putumayo	Villagarzon	1.06642	-76.62433	482
T. guianensis	*Myrmelachista*	MFT454	Putumayo	Villagarzon	1.06642	-76.62433	482
T. guianensis	*Myrmelachista*	MFT446	Putumayo	Villagarzon	1.06005	-76.62702	540
T. guianensis	*Myrmelachista*	MFT447	Putumayo	Villagarzon	1.06005	-76.62702	540
T. guianensis	*Myrmelachista*	MFT457	Putumayo	Villagarzon	1.05979	-76.62585	550
T. guianensis	*Azteca*	MFT427	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*NN*	MFT418	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT419	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT420	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT421	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT422	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT423	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT424	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT425	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Myrmelachista*	MFT426	Putumayo	Villagarzon	1.05531	-76.62315	562
T. guianensis	*Azteca*	MFT431	Putumayo	Villagarzon	1.05560	-76.62349	563
T. guianensis	*Azteca*	MFT428	Putumayo	Villagarzon	1.05560	-76.62349	563
T. guianensis	*Azteca*	MFT430	Putumayo	Villagarzon	1.05560	-76.62349	563

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*NN*	MFT432	Putumayo	Villagarzon	1.05560	-76.62349	563
T. guianensis	*NN*	MFT429	Putumayo	Villagarzon	1.05560	-76.62349	563
T. guianensis	*NN*	MFT468	Putumayo	Villagarzon	1.06591	-76.62154	564
T. guianensis	*NN*	MFT469	Putumayo	Villagarzon	1.06591	-76.62154	564
T. guianensis	*NN*	MFT470	Putumayo	Villagarzon	1.06665	-76.62077	579
T. guianensis	*Pheidole*	MFT473	Putumayo	Villagarzon	1.06665	-76.62077	579
T. guianensis	*NN*	MFT477	Putumayo	Villagarzon	1.06686	-76.62064	582
T. guianensis	*Azteca*	MFT438	Putumayo	Villagarzon	1.05577	-76.62359	588
T. guianensis	*Azteca*	MFT452	Putumayo	Villagarzon	1.06353	-76.62863	591
T. guianensis	*NN*	MFT449	Putumayo	Villagarzon	1.06353	-76.62863	591
T. guianensis	*NN*	MFT450	Putumayo	Villagarzon	1.06353	-76.62863	591
T. guianensis	*NN*	MFT451	Putumayo	Villagarzon	1.06353	-76.62863	591
T. guianensis	*NN*	MFT484	Putumayo	Villagarzon	1.06694	-76.62038	596
T. guianensis	*NN*	MFT485	Putumayo	Villagarzon	1.06694	-76.62038	596
T. guianensis	*NN*	MFT486	Putumayo	Villagarzon	1.06694	-76.62038	596
T. guianensis	*Pheidole*	MFT483	Putumayo	Villagarzon	1.06694	-76.62038	596
T. guianensis	*NN*	MFT478	Putumayo	Villagarzon	1.07123	-76.61575	718
T. guianensis	*NN*	MFT479	Putumayo	Villagarzon	1.07123	-76.61575	718
T. guianensis	*NN*	MFT480	Putumayo	Villagarzon	1.07123	-76.61575	718
T. guianensis	*NN*	MFT481	Putumayo	Villagarzon	1.07123	-76.61575	718
T. guianensis	*Myrmelachista*	MFT456	Putumayo	Villagarzon	1.06642	-76.62433	482
T. guianensis	*Azteca*	MFT463	Putumayo	Villagarzon	1.06509	-76.62222	529
T. guianensis	*NN*	MFT466	Putumayo	Villagarzon	1.06509	-76.62222	529
T. guianensis	*NN*	MFT467	Putumayo	Villagarzon	1.06509	-76.62222	529
T. guianensis	*Myrmelachista*	MFT445	Putumayo	Villagarzon	1.05942	-76.62523	548
T. guianensis	*NN*	MFT482	Putumayo	Villagarzon	1.06758	-76.61898	612
T. guianensis	*Azteca*	MFT490	Santander	Cimitarra	6.24815	-74.09312	163

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT489	Santander	Cimitarra	6.24817	-74.09305	165
T. guianensis	*Crematogaster*	MFT488	Santander	Cimitarra	6.24817	-74.09305	165
T. guianensis	*Pheidole*	MFT491	Santander	Cimitarra	6.24814	-74.09314	166
T. guianensis	*Pheidole*	MFT492	Santander	Cimitarra	6.24804	-74.09318	167
T. guianensis	*Azteca*	MFT495	Santander	Cimitarra	6.24787	-74.09330	171
T. guianensis	*Azteca*	MFT493	Santander	Cimitarra	6.24791	-74.09324	173
T. guianensis	*Azteca*	MFT494	Santander	Cimitarra	6.24789	-74.09326	173
T. guianensis	*Azteca*	MFT596	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT597	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT598	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT600	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT601	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT602	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT603	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT592	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Pheidole*	MFT593	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Pheidole*	MFT594	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Pheidole*	MFT595	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Pheidole*	MFT599	Santander	Barrancabermeja	7.07248	-73.74724	92
T. guianensis	*Azteca*	MFT134	Valle del Cauca	Buenaventura	4.008196	-77.393975	17
T. guianensis	*Azteca*	MFT135	Valle del Cauca	Buenaventura	4.008196	-77.393975	17
T. guianensis	*Azteca*	MFT136	Valle del Cauca	Buenaventura	4.008196	-77.393975	17
T. guianensis	*Azteca*	MFT137	Valle del Cauca	Buenaventura	4.008196	-77.393975	17
T. guianensis	*Azteca*	MFT138	Valle del Cauca	Buenaventura	4.008196	-77.393975	17
T. guianensis	*NN*	MFT139	Valle del Cauca	Buenaventura	4.008196	-77.393975	17
T. guianensis	*Azteca*	MFT327	Valle del Cauca	Buenaventura	3.95731	-77.01983	27
T. guianensis	*Azteca*	MFT328	Valle del Cauca	Buenaventura	3.95731	-77.01983	27

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT330	Valle del Cauca	Buenaventura	3.95731	-77.01983	27
T. guianensis	*Azteca*	MFT333	Valle del Cauca	Buenaventura	3.96816	-77.00461	33
T. guianensis	*Azteca*	MFT334	Valle del Cauca	Buenaventura	3.96816	-77.00461	33
T. guianensis	*Azteca*	MFT335	Valle del Cauca	Buenaventura	3.96816	-77.00461	33
T. guianensis	*Azteca*	MFT336	Valle del Cauca	Buenaventura	3.96816	-77.00461	33
T. guianensis	*Azteca*	MFT337	Valle del Cauca	Buenaventura	3.96816	-77.00461	33
T. guianensis	*Azteca*	MFT321	Valle del Cauca	Buenaventura	3.95806	-77.01912	42
T. guianensis	*Azteca*	MFT314	Valle del Cauca	Buenaventura	3.95839	-77.01857	47
T. guianensis	*Azteca*	MFT315	Valle del Cauca	Buenaventura	3.95839	-77.01857	47
T. guianensis	*Azteca*	MFT316	Valle del Cauca	Buenaventura	3.95839	-77.01857	47
T. guianensis	*Azteca*	MFT317	Valle del Cauca	Buenaventura	3.95823	-77.01886	47
T. guianensis	*Azteca*	MFT318	Valle del Cauca	Buenaventura	3.95823	-77.01886	47
T. guianensis	*Azteca*	MFT319	Valle del Cauca	Buenaventura	3.95823	-77.01886	47
T. guianensis	*Azteca*	MFT332	Valle del Cauca	Buenaventura	3.96765	-77.00463	47
T. guianensis	*NN*	MFT320	Valle del Cauca	Buenaventura	3.95823	-77.01886	47
T. guianensis	*Azteca*	MFT353	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT354	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT355	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT356	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT357	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*NN*	MFT307	Valle del Cauca	Buenaventura	3.95948	-77.01772	50
T. guianensis	*Azteca*	MFT312	Valle del Cauca	Buenaventura	3.95891	-77.01796	59
T. guianensis	*Azteca*	MFT367	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
T. guianensis	*Azteca*	MFT368	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
T. guianensis	*Azteca*	MFT365	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
T. guianensis	*Azteca*	MFT366	Valle del Cauca	Buenaventura	3.95360	-76.99190	64
T. guianensis	*Azteca*	MFT369	Valle del Cauca	Buenaventura	3.95360	-76.99190	64

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*NN*	MFT133	Valle del Cauca	Buenaventura	3.828908	-76.882435	96
T. guianensis	*Azteca*	MFT360	Valle del Cauca	Buenaventura	3.95281	-76.99081	103
T. guianensis	*Azteca*	MFT361	Valle del Cauca	Buenaventura	3.95281	-76.99081	103
T. guianensis	*NN*	MFT359	Valle del Cauca	Buenaventura	3.95281	-76.99081	103
T. guianensis	*NN*	MFT362	Valle del Cauca	Buenaventura	3.95281	-76.99081	103
T. guianensis	*NN*	MFT363	Valle del Cauca	Buenaventura	3.95281	-76.99081	103
T. guianensis	*Azteca*	MFT126	Valle del Cauca	Buenaventura	3.830342	-76.886715	125
T. guianensis	*Azteca*	MFT127	Valle del Cauca	Buenaventura	3.830342	-76.886715	125
T. guianensis	*Azteca*	MFT128	Valle del Cauca	Buenaventura	3.830342	-76.886715	125
T. guianensis	*Azteca*	MFT130	Valle del Cauca	Buenaventura	3.830342	-76.886715	125
T. guianensis	*Azteca*	MFT125	Valle del Cauca	Buenaventura	3.830342	-76.886715	125
T. guianensis	*NN*	MFT129	Valle del Cauca	Buenaventura	3.830342	-76.886715	125
T. guianensis	*Azteca*	MFT302	Valle del Cauca	Buenaventura	3.95982	-77.01752	48
T. guianensis	*Azteca*	MFT303	Valle del Cauca	Buenaventura	3.95982	-77.01752	48
T. guianensis	*Azteca*	MFT304	Valle del Cauca	Buenaventura	3.95982	-77.01752	48
T. guianensis	*Azteca*	MFT305	Valle del Cauca	Buenaventura	3.95982	-77.01752	48
T. guianensis	*Azteca*	MFT340	Valle del Cauca	Buenaventura	3.97001	-77.00346	48
T. guianensis	*Azteca*	MFT346	Valle del Cauca	Buenaventura	3.97001	-77.00346	48
T. guianensis	*Azteca*	MFT306	Valle del Cauca	Buenaventura	3.95982	-77.01752	48
T. guianensis	*Azteca*	MFT338	Valle del Cauca	Buenaventura	3.97001	-77.00346	48
T. guianensis	*Azteca*	MFT339	Valle del Cauca	Buenaventura	3.97001	-77.00346	48
T. guianensis	*Azteca*	MFT345	Valle del Cauca	Buenaventura	3.97001	-77.00346	48
T. guianensis	*Azteca*	MFT348	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT349	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT350	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT351	Valle del Cauca	Buenaventura	3.97306	-77.00214	49
T. guianensis	*Azteca*	MFT352	Valle del Cauca	Buenaventura	3.97306	-77.00214	49

Table B.1 continued from previous page

Plant	Ant	Collection code	Department	Locality	Latitude	Longitude	Altitude (m)
T. guianensis	*Azteca*	MFT308	Valle del Cauca	Buenaventura	3.95948	-77.01772	50
T. guianensis	*Azteca*	MFT309	Valle del Cauca	Buenaventura	3.95948	-77.01772	50
T. guianensis	*NN*	MFT310	Valle del Cauca	Buenaventura	3.95948	-77.01772	50
T. guianensis	*Azteca*	MFT311	Valle del Cauca	Buenaventura	3.95966	-77.01755	61
T. guianensis	*Azteca*	MFT342	Valle del Cauca	Buenaventura	3.97011	-77.00326	66
T. guianensis	*Azteca*	MFT343	Valle del Cauca	Buenaventura	3.97011	-77.00326	66
T. guianensis	*Azteca*	MFT344	Valle del Cauca	Buenaventura	3.97011	-77.00326	66
T. macrophysca	*Azteca*	MFT157	Amazonas	Leticia	-4.120414	-69.954895	100
T. spadiciflora	*Pheidole*	MFT224	Choco	Arusi	5.57331	-77.50000	66
T. spadiciflora	*Pheidole*	MFT189	Choco	Arusi	5.57165	-77.50081	106
T. spadiciflora	*Pheidole*	MFT190	Choco	Arusi	5.57165	-77.50081	106
T. spadiciflora	*Pheidole*	MFT191	Choco	Arusi	5.57165	-77.50081	106
T. spadiciflora	*Pheidole*	MFT192	Choco	Arusi	5.57165	-77.50081	106
T. spadiciflora	*Pheidole*	MFT253	Choco	Arusi	5.57518	-77.49706	106
T. spadiciflora	*Pheidole*	MFT254	Choco	Arusi	5.57518	-77.49706	106
T. spadiciflora	*Pheidole*	MFT256	Choco	Arusi	5.57532	-77.49722	107
T. spadiciflora	*Pheidole*	MFT257	Choco	Arusi	5.57532	-77.49722	107
T. spadiciflora	*Pheidole*	MFT260	Choco	Arusi	5.57599	-77.50022	116
T. spadiciflora	*Pheidole*	MFT212	Choco	Arusi	5.57136	-77.50111	53
T. spadiciflora	*Pheidole*	MFT204	Choco	Arusi	5.57137	-77.50076	64
T. spadiciflora	*Pheidole*	MFT241	Choco	Arusi	5.57160	-77.49809	65
T. spadiciflora	*Pheidole*	MFT242	Choco	Arusi	5.57160	-77.49809	65
T. spadiciflora	*Pheidole*	MFT233	Choco	Arusi	5.57283	-77.49967	71
T. spadiciflora	*Pheidole*	MFT235	Choco	Arusi	5.57283	-77.49967	71
T. spadiciflora	*Pheidole*	MFT132	Valle del Cauca	Buenaventura	3.828908	-76.882435	96
T. stenoptera	*Azteca*	MFT293	Meta	S.J. de Arama	3.273785	-73.893999	362
T. stenoptera	*Pheidole*	MFT294	Meta	S.J. de Arama	3.273785	-73.893999	362

Table B.2: Melastomataceae plants and their associated ants collected from 2013 to 2016 by location and sublocation. Areas correspond to the region relative to the Andean cordillera

Area	Department	Location	*T. guanensis*	Other Melastomataceae	Total samples
West north	Choco	Arusi 1	9	7	16
	Choco	Arusi 2	27	13	40
	Choco	Arusi 3	15	3	18
	Choco	Arusi 4	10	5	15
	Valle del Cauca	San Cipriano	12	0	12
	Valle del Cauca	Buenaventura	6	0	6
	Valle del Cauca	Bajo Calima	23	8	31
	Valle del Cauca	Bajo Calima	35	8	43
	Valle del Cauca	Bajo Calima	10	6	16
Central north	Antioquia	San Luis	16	0	16
	Antioquia	San Carlos	10	2	12
	Antioquia	Amalfi 1	24	0	24
	Antioquia	Amalfi 2	19	0	19
	Santander	La India	8	0	8
	Santander	Barrancabermeja*	12	0	12
East north	Casanare	Villanueva*	9	0	9
	Casanare	Tauramena 1	8	0	8
	Casanare	Tauramena 2	8	0	0
	Meta	S.J. de Arama 1	15	1	16
	Meta	S.J. de Arama 2	12	0	12
	Meta	S.J. de Arama 3	12	0	12
	Meta	Villavicencio 1	5	0	5
	Meta	Acacias	7	0	7
	Meta	Villavicencio 2	9	0	9
Amazon west	Putumayo	Villagarzon 1	20	24	44

Table B.2 continued from previous page

	Putumayo	Villagarzon 2	13	13	26
	Putumayo	Villagarzon 3	17	10	27
Amazon east	Amazonas	Kilometros	19	12	31
	Amazonas	Tarapaca	20	0	20
	Amazonas	Puerto Nariño	10	0	10
Other	Tolima and Cundinamarca		0	7	7
Total			420	119	531

*Samples collected during 2016

Table B.3: Ants accessions for each genera collected from *T. guianensis* and other myrmecophyte Melastomataceae plants in each collecting site around Colombia.

Locality	Ant genus	Host plant	
		T. guianensis	**Other Melastomataceae***
Amazonas	*Allomerus*	4	1
	Azteca	16	1
	Camponotus	-	1
	Crematogaster	-	1
	Nylanderia	-	1
	Pheidole	3	2
Antioquia	*Azteca*	54	1
	Crematogaster	2	-
	Pheidole	2	1
	Wasmannia	2	-
Casanare	*Azteca*	19	-
	Camponotus	1	-
	Pheidole	4	-
Choco	*Azteca*	-	28
	Pheidole	41	40
Meta	*Tapinoma*	3	-
	Azteca	46	1
	Myrmelachista	1	-
	Pheidole	1	1
	Solenopsis	2	-
Putumayo	*Allomerus*	-	3
	Azteca	28	26
	Crematogaster	-	1
	Myrmelachista	14	5

Table B.3 continued from previous page

	Pheidole	2	6
Santander	*Azteca*	17	-
	Crematogaster	1	-
	Pheidole	6	-
Valle del Cauca	*Azteca*	38	-
	Pheidole	7	-
	Wasmannia	-	1

**T. macrophylla, T. bullifera, T. stenoptera, T. spadiciflora, Miconia sp., Maieta sp., Conostegia sp., Cecropia sp., Henriettella sp.*

Table B.4: List of NCBI accession numbers of the sequences used in phylogenetic analyses

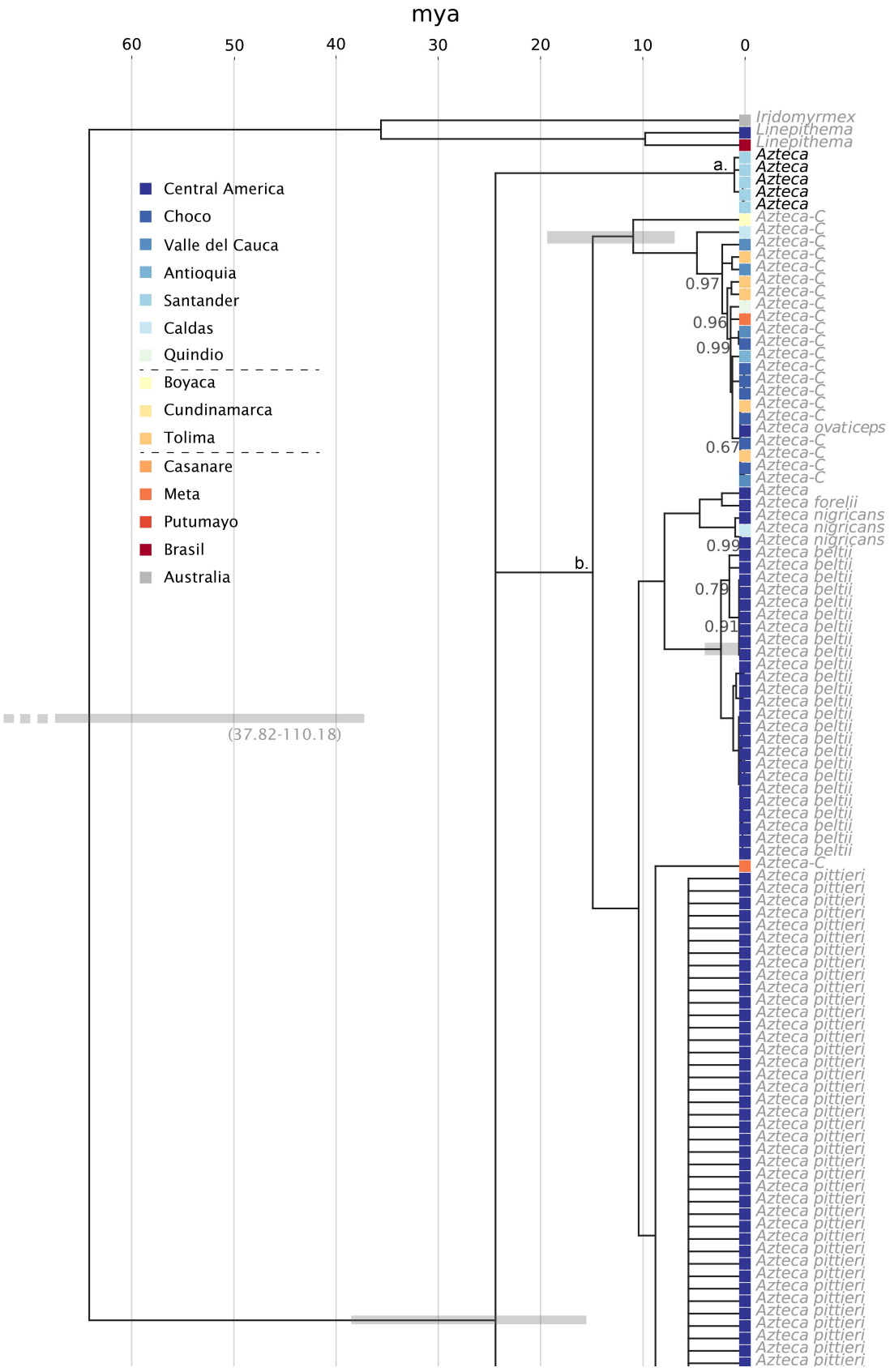
Species	Accession number	Species	Accession number
Azteca beltii	JQ867700	*Azteca nigricans*	JQ867868
Azteca beltii	JQ867701	*Azteca nigricans*	JQ867869
Azteca beltii	JQ867706	*Azteca ovaticeps*	JQ867726
Azteca beltii	JQ867709	*Azteca pittieri*	JQ867805
Azteca beltii	JQ867714	*Azteca pittieri*	JQ867809
Azteca beltii	JQ867746	*Azteca pittieri*	JQ867810
Azteca beltii	JQ867767	*Azteca pittieri*	JQ867811
Azteca beltii	JQ867785	*Azteca pittieri*	JQ867812
Azteca beltii	JQ867789	*Azteca pittieri*	JQ867814
Azteca beltii	JQ867813	*Azteca pittieri*	JQ867815
Azteca beltii	JQ867832	*Azteca pittieri*	JQ867816
Azteca beltii	JQ867863	*Azteca pittieri*	JQ867817
Azteca beltii	JQ867864	*Azteca pittieri*	JQ867818
Azteca beltii	JQ867865	*Azteca pittieri*	JQ867819
Azteca beltii	JQ867866	*Iridomyrmex purpureus*	FJ161760
Azteca forelii	JQ867787	*Linepithema iniquum*	FJ161780
Azteca nigricans	JQ867693	*Linepithema micans*	FJ161787

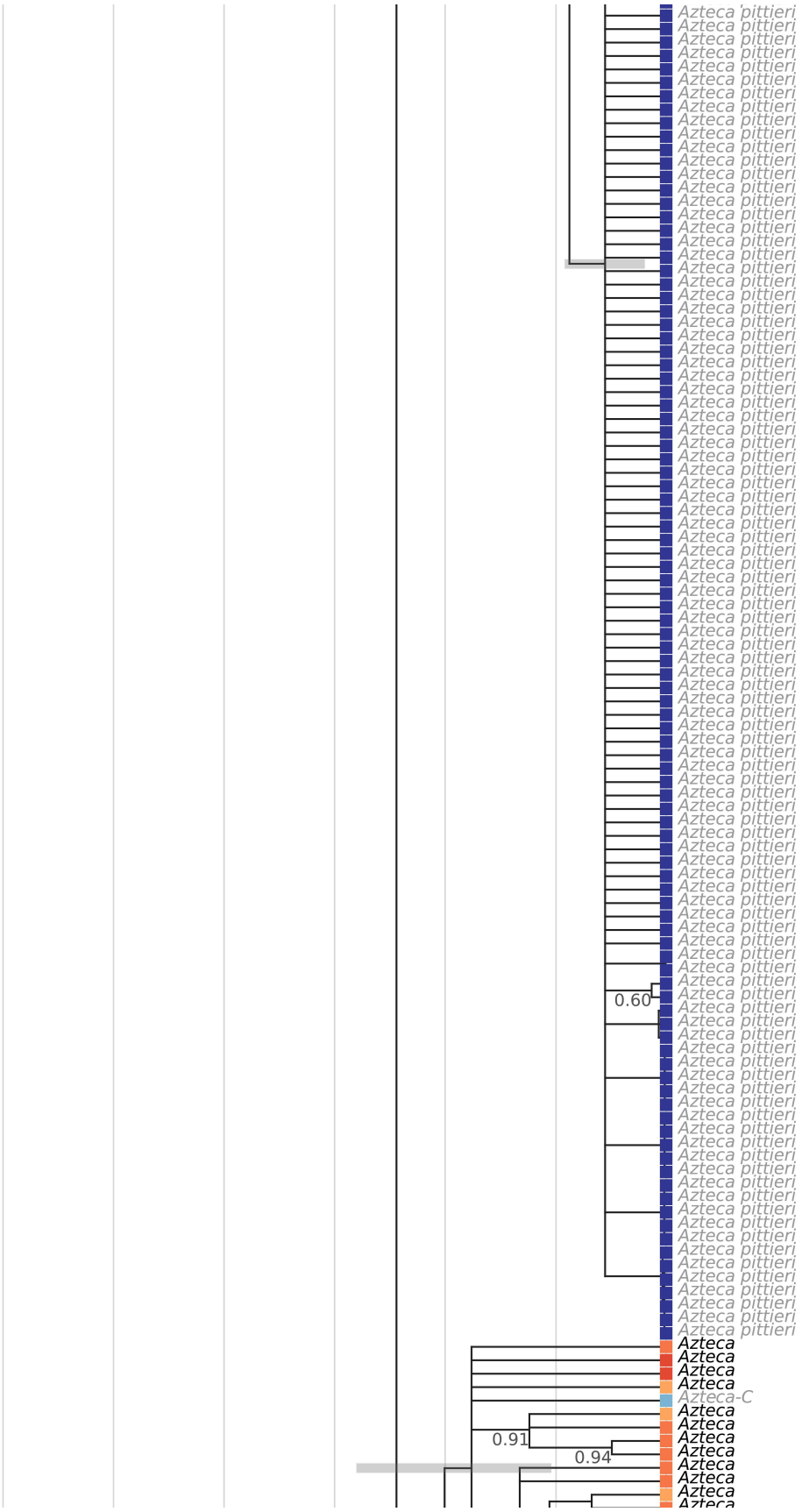
Table B.5: Number of MOTUs delimited using **jMOTU** using different percentage of sequence length and pair bases cut-offs

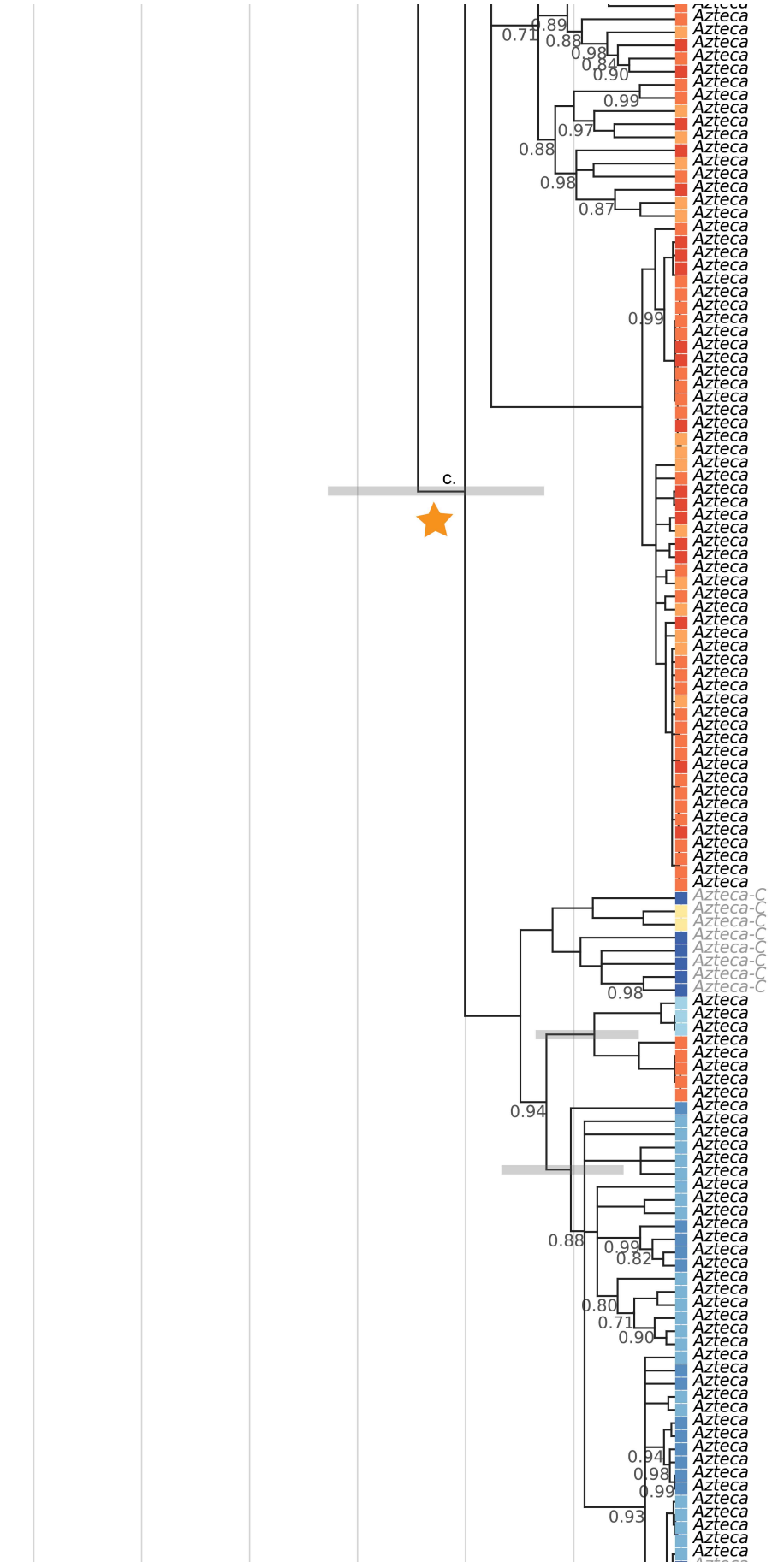
COI			**ITS2**		
Number of clusters	**Threshold value in % sequence length**	**Cut-off value (bp)**	**Number of clusters**	**Threshold value in % sequence length**	**Cut-off value (bp)**
109	0.15	1	27	0.25	2
51	0.46	3	14	0.62	5
21	1.52	10	13	1	8
17	2.28	15	12	1.25	10
12	2.88	19	11	1.87	15
11	3.04	20	10	2.87	23
9	4.1	27	7	3	24
8	4.4	29	6	3.37	27
8	6.07	40	4	4.37	35
7	7.13	47	3	4.99	40
-	-	-	2	6.87	55
-	-	-	1	9.36	75

Table B.6: Number of MOTUs delimited using **ABGD** using different percentage of sequence length and pair bases cut-offs

COI		**ITS2**	
Number of clusters	**Threshold value (JC69)**	**Number of clusters**	**Threshold value (JC69)**
18	0.00623	12	0.00128
15	0.00933	11	0.00258
13	0.0125	-	-


mya
60
50
40
30
20
10
0
Central America
Choco
Valle del Cauca
Antioquia
Santander
Caldas
Quindio
Boyaca
Cundinamarca
Tolima
Casanare
Meta
Putumayo
Brasil
Australia
(37.82-110.18)
a.
b.
0.97
0.96
0.99
0.67
0.99
0.79
0.91
Iridomyrmex
Linepithema
Linepithema
Azteca
Azteca
Azteca
Azteca
Azteca
Azteca-C
Azteca ovaticeps
Azteca forelii
Azteca nigricans
Azteca nigricans
Azteca nigricans
Azteca beltii
Azteca pittieri


Azteca pittieri
0.60
Azteca
Azteca-C
0.91
0.94


c.
Azteca
Azteca-C
0.71
0.89
0.88
0.98
0.84
0.90
0.99
0.97
0.88
0.98
0.87
0.99
0.98
0.94
0.88
0.99
0.82
0.80
0.71
0.90
0.94
0.98
0.99
0.93

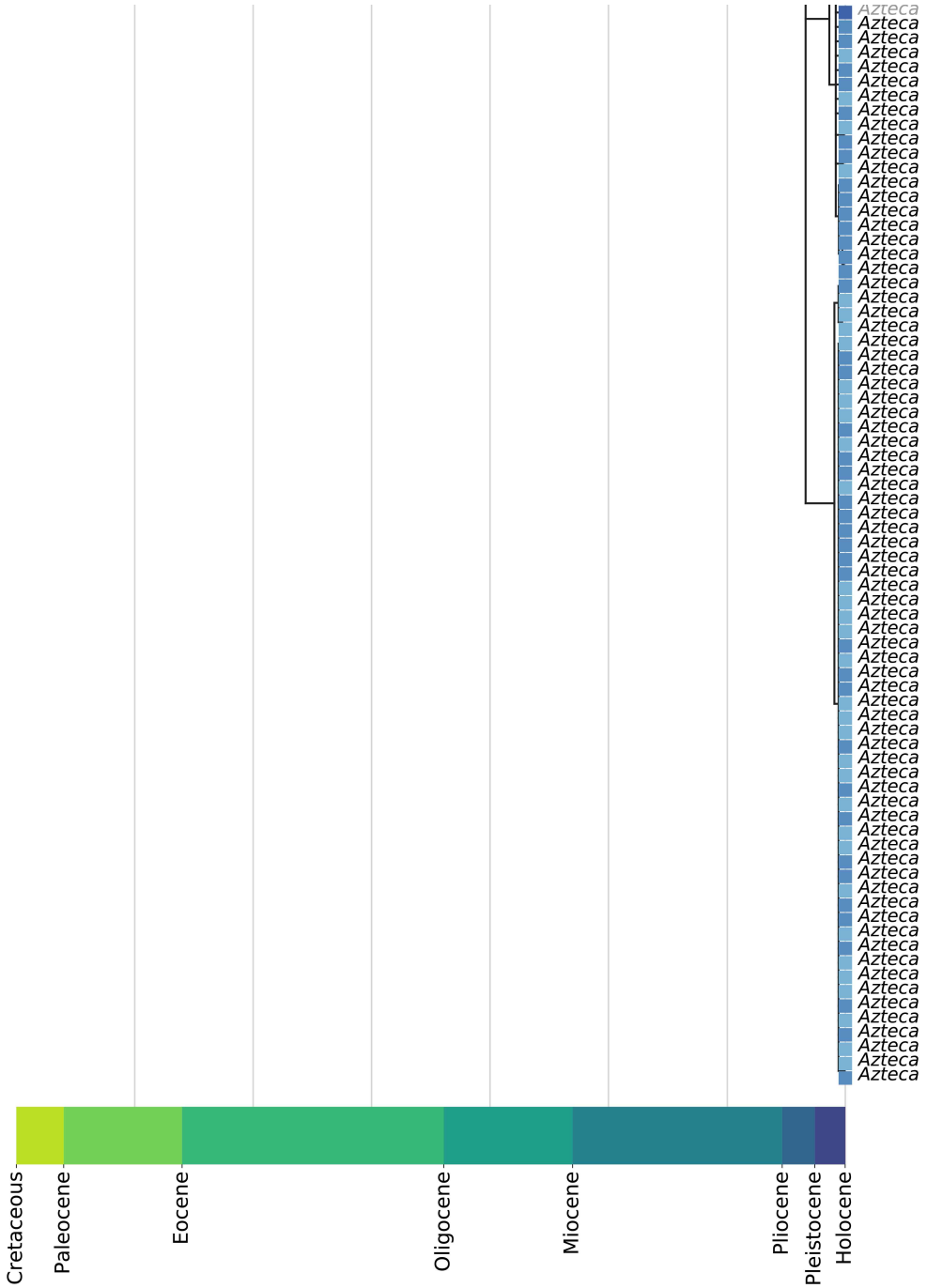

Azteca
Cretaceous
Paleocene
Eocene
Oligocene
Miocene
Pliocene
Pleistocene
Holocene

Figure B.1: Time calibrated phylogeny of ITS2 *Azteca* sequences including all *Azteca* sequences available as opposed to a subsample of sequences per species. The calibration was estimated using a birth-death model (Gernhard, 2008). The tMRCA of all *Azteca* was calibrated using a Dominican amber fossil of *Azteca* dated 20-15 million years old, providing a minimum age estimate for the genus (Wilson, 1985). A lognormal distributed prior was used with a mean (in real space) of 20 Mya, offset of 15 Mya, and a standard deviation of 1 so the 95% interval ranges from 15-75 Mya and includes estimates for *Azteca* tMRCAs in Moreau et al. (2006) and Ward et al. (2010). A strict clock model was set with a mean rate 1.84% per My as reported by Papadopoulou et al. (2010) and Ho and Lo (2013). Sample names in black were collected during this study, those in grey are NCBI sequences or *Cecropia*-associated *Azteca*. Posterior probabilities between 0.5 and 0.99 are shown. Branches with posterior probabilities lower than 0.5 are collapsed.

APPENDIX

C

APPENDIX CHAPTER 3

Table C.1: NCBI accession numbers, *Azteca* assemblies, and percentage of single-copy completed, duplicated, fragmented and missing **BUSCO** genes from a database of n=148 orthologue genes.

Genome	Complete	Duplicated	Fragmented	Missing
GCA_000167475 *wDrosophila ananassae*	55.4	1.4	6.8	36.4
GCA_000174095 *wMuscidifurax uniraptor*	50.7	0	3.4	45.9
GCA_000306885 *wOnchocerca ochengi*	78.4	0	2	19.6
GCA_001266585 *wOperophtera brumata*	83.1	0	0.7	16.2
GCF_000008025 *wDrosophila melanogaster*	87.2	0	2	10.8
GCF_000008385 *wBrugia malayi*	85.1	0	0.7	14.2
GCF_000022285 *Wolbachia sp.* wRi	87.2	0	2	10.8
GCF_000073005 *wCulex quinquefasciatus*	87.2	0	1.4	11.4
GCF_000204545 *wNasonia vitripennis*	85.1	0	2	12.9
GCF_000208785 *wCulex pipiens molestus*	83.8	0	2	14.2
GCF_000331595 *wDiaphorina citri*	87.2	0	1.4	11.4
GCF_000333795 *wDrosophila suzukii*	87.8	0	1.4	10.8
GCF_000376585 *wDrosophila simulans*	87.2	0	2	10.8
GCF_000530755 *wOncocerca volvulus_Cameroon*	73	0	2	25
GCF_000689175 *wGlossina morsitan morsitans*	72.3	0.7	4.7	22.3
GCF_000742435 *Wolbachia pipientis*	85.8	0	1.4	12.8
GCF_000829315 *wCimex lectularius*	85.1	0	4.7	10.2
GCF_001439985 *wTrichogramma pretiosum*	83.8	0	2	14.2
GCF_001637495 *wLaodelphax striatella*	87.8	0	1.4	10.8
GCF_001648015 *wDactylopius coccus*	75	8.8	4.1	12.1
GCF_001675695 *wNomada flava*	87.2	0.7	1.4	10.7
GCF_001675715 *wNomada leucophthalma*	87.8	0	1.4	10.8
GCF_001675775 *wNomada panzeri*	85.1	2	1.4	11.5
GCF_001675785 *wNomada ferruginata*	85.8	0.7	2.7	10.8
GCF_001758565 *wDrosophila incompta*	81.1	0	3.4	15.5
MFT327	80.4	0	4.1	15.5

Table C.1 continued from previous page

MFT334	77.7	0.7	4.1	17.5
MFT400	83.8	0	3.4	12.8
MFT493	75.7	0	3.4	20.9
MFT591	33.1	0.7	16.2	50

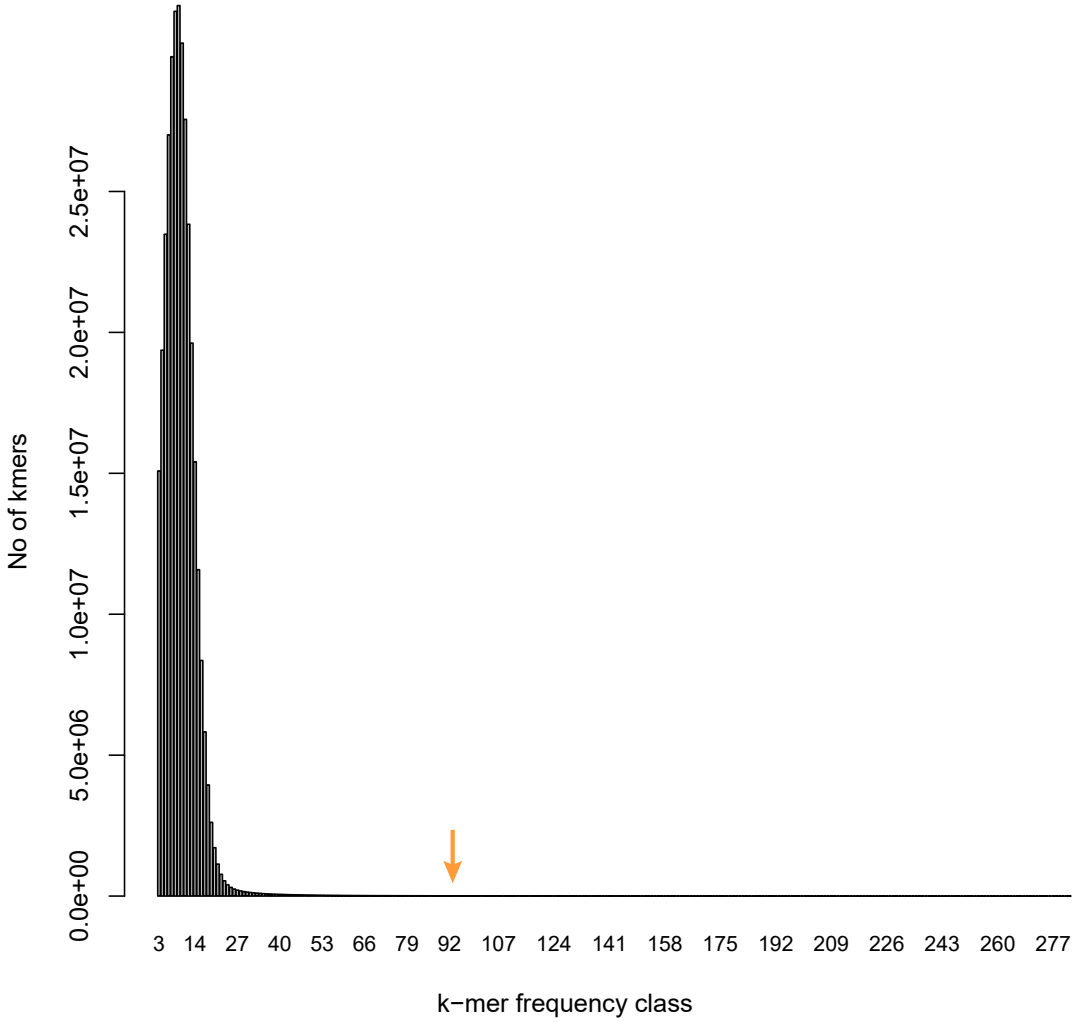


Figure C.1: Kmer frequency plot of the MFT151 *Azteca* assembly. Kmers present in the assembly more than 70 times can be considered as signal for repetitive regions (right tail of the distribution).

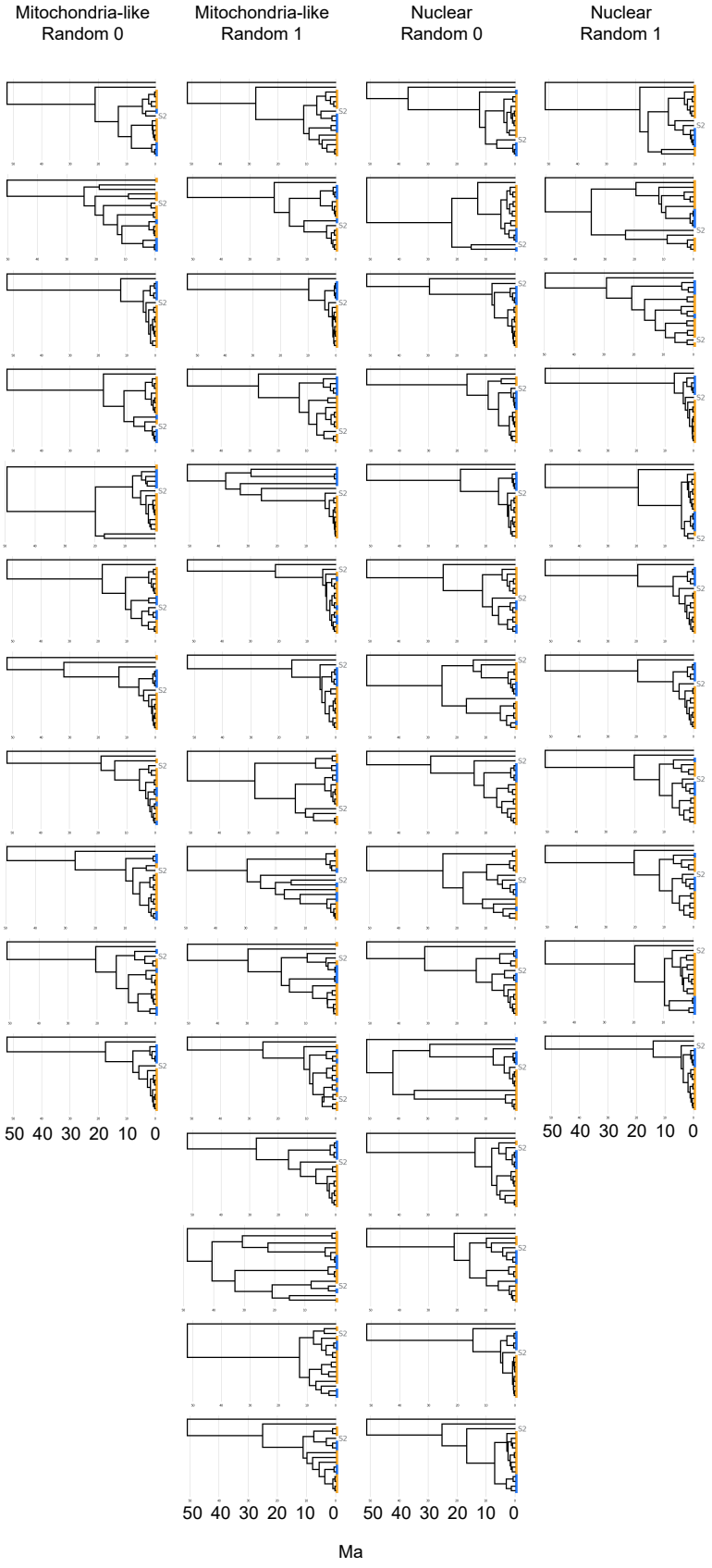


Figure C.2: Alternative gene tree topologies showing the position of S2 with respect to Eastern *Azteca* (orange) and Western *Azteca* (blue).

APPENDIX

D

APPENDIX CHAPTER 4

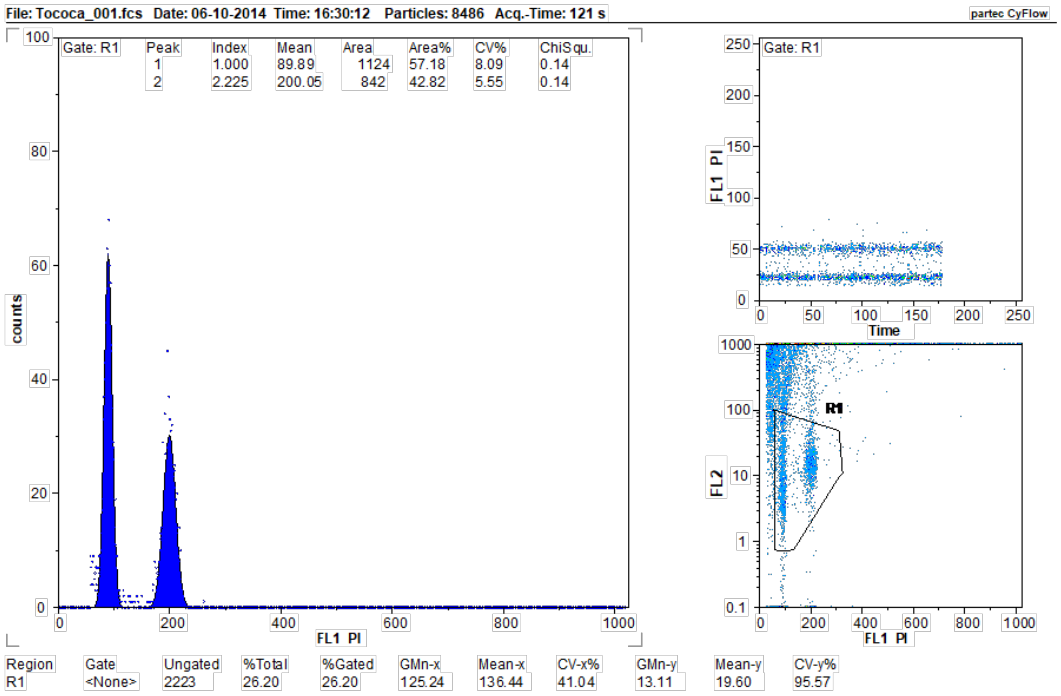


Figure D.1: DNA histograms from the flow cytometry analysis of one (out of three) *T. guianensis* plants in the living collection of the Munich Botanic Garden. The leaf tissue was kindly donated by the collection.

Table D.1: List of NCBI accession numbers of the sequences used in phylogenetic analyses.

Marker	Species	Accession number
ndhF	*Clidemia capillaris*	EU056121
	Clidemia dentata	EU002211
	Clidemia rubra	AF215579
	Maieta guianensis	AF215581
	Miconia brasiliensis	GQ139322
	Miconia pyramidalis	JF831979
	Miconia superba	EU056105
	Phainantha laxiflora	JF831980
	Phainantha shuariorum	JF831981
	Tococa broadwayi	EU056134
	Tococa caquetana	EU056135
	Tococa guianensis	EU056136
	Tococa perclara	EU056137
	Tococa platyphylla	EU056138
	Tococa spadiciflora	EU056139
ITS	*Miconia cinerea*	KJ418737
	Miconia ferruginea	AY460510
	Tococa bolivarensis	AY460547
	Tococa broadwayi	AY460548
	Tococa capitata	AY460549
	Tococa caquetana	AY460550

Table D.1 continued from previous page

Tococa caudata	AY460551
Tococa coronata	AY460552
Tococa discolor	EU055895
Tococa gonoptera	AY460553
Tococa guianensis	AY460554
Tococa macrophysca	AY460555
Tococa macrosperma	AY460556
Tococa nitens	AY460557
Tococa perclara	AY460558
Tococa platyphylla	EU055896
Tococa quadrialata	EF418922
Tococa raggiana	AY460559
Tococa rotundifolia	AY460560
Tococa spadiciflora	EU055897
Tococa subciliata	AY460561

Table D.2: Length, nucleotide diversity, segregating sites and Watterson's Theta values for all the single copy **BUSCO** alignments and for the quartile groups of genes.

		Range	**Mean**	**Median**
All single copy BUSCO	Alignment length	213 - 16693	1850.56	1391.88
	Nucleotide diversity (pi)	0-0.44	0.03	0.05
	No. of segregating sites	0-865	70.66	115.57
	Watterson's theta	0-471.82	38.54	63.04
Quartile 1 Low diversity	Alignment length	213 - 2976	1055.58	954
	Nucleotide diversity (pi)	0 - 0.036	0.0076	0.0067
	No. of segregating sites	0 - 16	10.47	11
	Watterson's theta	0 - 8.72	5.71	6
Quartile 2 Medium diversity	Alignment length	350 - 6734	1671.33	1423.5
	Nucleotide diversity (pi)	0.0045 - 0.042	0.013	0.011
	No. of segregating sites	18 - 34	24.96	24.5
	Watterson's theta	9.81 - 18.54	13.61	13.36
Quartile 3 High diversity	Alignment length	304 - 6911	2224.23	1950.5
	Nucleotide diversity (pi)	0.0054 - 0.12	0.020	0.015
	No. of segregating sites	36 - 66	47.63	46
	Watterson's theta	19.63 - 36.0	25.98	25.09
Quartile 4 Super-high diversity	Alignment length	473 - 16693	2413.03	1953
	Nucleotide diversity (pi)	0.0039 - 0.439	0.081	0.049
	No. of segregating sites	68 - 865	199.39	120
	Watterson's theta	37.09 - 471	108.76	65.45

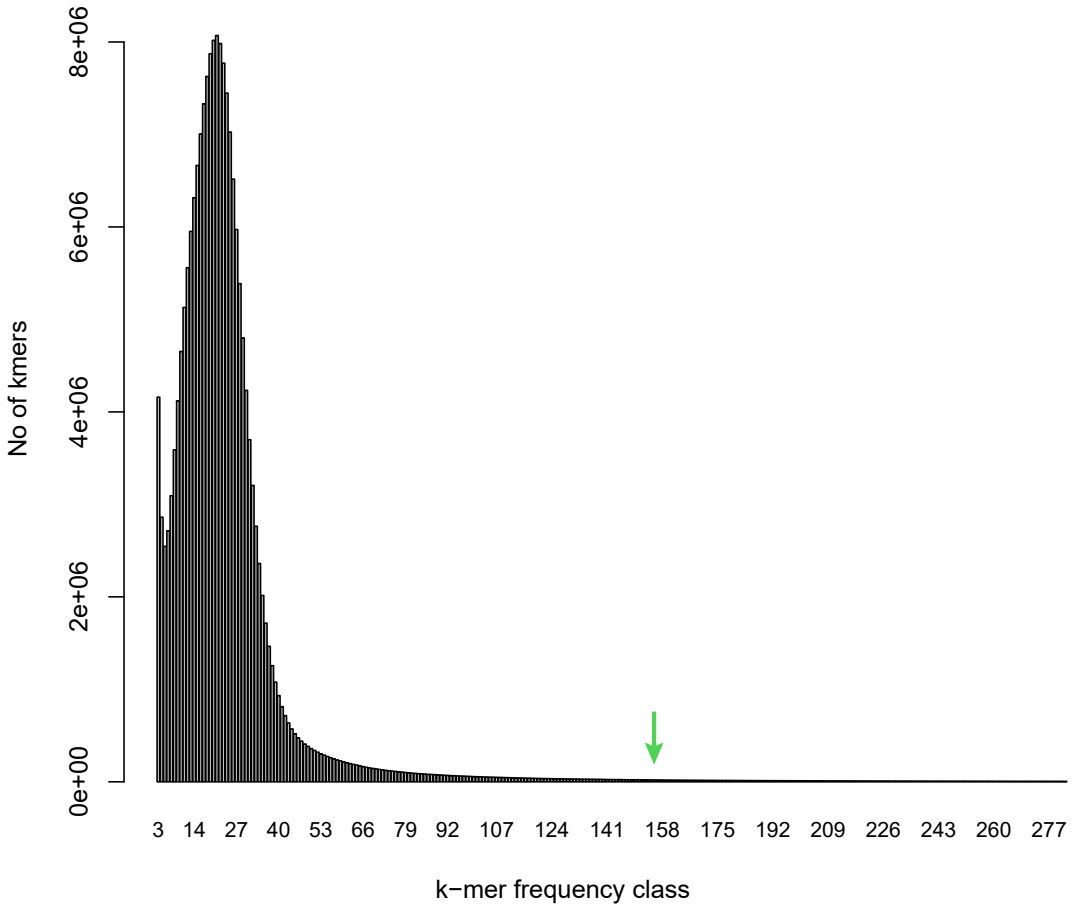


Figure D.2: Kmer frequency plot of the PMFT244 *Tococa* assembly. Kmers present in the assembly more than 70 times can be considered as signal for repetitive regions (right tail of the distribution).

Table D.3: Name, length, % of identity and reference species of the best **Blast** hit result for the **BUSCO** single-copy genes selected from quartile two.

NCBI accession	Length	e-value	% identity	Protein	Species
XP_010025000.1	459	0	73.203	PREDICTED: uncharacterized protein	*Eucalyptus grandis*
XP_010025158.1	402	1.41E-141	55.473	PREDICTED: putative uncharacterized protein	*Eucalyptus grandis*
KCW79033.1	1300	0	56.308	hypothetical protein	*Eucalyptus grandis*
KCW48379.1	384	0	79.948	hypothetical protein	*Eucalyptus grandis*
EOY01717.1	236	5.67E-95	66.525	Chaperone DnaJ-domain superfamily protein	*Theobroma cacao*
XP_002518362.1	364	0	76.374	zinc binding dehydrogenase	*Ricinus communis*
KCW57362.1	667	0	63.118	hypothetical protein	*Eucalyptus grandis*
XP_010062188.1	679	0	62.887	PREDICTED: putative pentatricopeptide repeat containing protein At1g77010, mitochondrial	*Eucalyptus grandis*
XP_009779924.1	985	0	39.289	PREDICTED: disease resistance protein RPS2-like	*Nicotiana sylvestris*
KCW64398.1	229	5.61E-98	67.686	hypothetical protein EUGRSUZ_G02016	*Eucalyptus grandis*
XP_004135707.1	123	4.37E-44	59.35	PREDICTED: uncharacterized protein At4g29660	*Cucumis sativus*
KDO71828.1	246	4.08E-86	58.537	hypothetical protein	*Citrus sinensis*
XP_010055087.1	184	2.99E-109	80.435	PREDICTED: uncharacterized protein LOC104443413 isoform X2	*Eucalyptus grandis*

Table D.3 continued from previous page

KCW51183.1	446	0	94.17	hypothetical protein	*Eucalyptus grandis*
XP_006438159.1	227	4.12E-97	65.198	hypothetical protein CICLE_v10032228mg	*Citrus clementina*
XP_010038345.1	130	4.28E-45	60.769	PREDICTED: zinc finger HIT domain containing protein 3 isoform X2	*Eucalyptus grandis*
KCW85363.1	571	0	66.55	hypothetical protein EUGRSUZ_B02194	*Eucalyptus grandis*
XP_010033178.1	303	9.46E-159	69.637	PREDICTED: nudix hydrolase 9 isoform X1	*Eucalyptus grandis*
XP_010047728.1	419	0	82.1	PREDICTED: uncharacterized protein LOC104436595 isoform X1	*Eucalyptus grandis*
XP_010038091.1	232	1.01E-114	72.414	PREDICTED: fatty-acid-binding protein 1	*Eucalyptus grandis*
XP_010055475.1	355	1.38E-167	66.761	PREDICTED: uncharacterized protein	*Eucalyptus grandis*
XP_010040693.1	343	3.12E-153	66.181	PREDICTED: probable galactose-1-phosphate uridyltransferase	*Eucalyptus grandis*
XP_010058980.1	200	7.96E-109	80.5	PREDICTED: 33 kDa ribonucleoprotein, chloroplastic	*Eucalyptus grandis*
XP_010048386.1	542	0	70.111	PREDICTED: pentatricopeptide repeat containing protein At5g10690	*Eucalyptus grandis*
XP_010056324.1	268	7.72E-140	71.269	PREDICTED: uncharacterized protein LOC104444365	*Eucalyptus grandis*
XP_011008938.1	391	0	65.217	PREDICTED: uncharacterized protein LOC105114170	*Populus euphratica*
XP_010062061.1	257	1.11E-153	82.879	PREDICTED: probable F-actin-capping protein subunit beta	*Eucalyptus grandis*

Table D.3 continued from previous page

EOY25237.1	584	0	62.842	Tetratricopeptide repeat-like superfamily protein	*Theobroma cacao*
XP_010061464.1	379	0	85.752	PREDICTED: cysteine synthase 2 isoform X1	*Eucalyptus grandis*
XP_010097667.1	243	1.55E-145	81.893	50S ribosomal protein L3-1	*Morus notabilis*
XP_010031740.1	292	5.49E-150	70.548	PREDICTED: phytochromobilin ferredoxin oxidoreductase, chloroplastic	*Eucalyptus grandis*
XP_010023588.1	563	0	82.06	PREDICTED: 15-cis-phytoene desaturase, chloroplastic chromoplastic isoform X1	*Eucalyptus grandis*
EOY10143.1	290	1.17E-73	47.586	Plasma membrane isoform 1	*Theobroma cacao*
CDP05963.1	239	2.61E-81	61.506	unnamed protein product	*Coffea canephora*
XP_010027529.1	649	0	76.579	PREDICTED: uncharacterized protein LOC104418018	*Eucalyptus grandis*
EOY09230.1	214	2.88E-78	53.738	N-terminal glutamine amidohydrolase isoform 1	*Theobroma cacao*
XP_010037073.1	522	0	56.322	PREDICTED: uncharacterized protein LOC104425902 isoform X1	*Eucalyptus grandis*
XP_010050654.1	124	9.35E-63	73.387	PREDICTED: uncharacterized protein LOC104439249 isoform X3	*Eucalyptus grandis*
XP_010038067.1	420	2.07E-155	56.905	PREDICTED: uncharacterized protein LOC104426639	*Eucalyptus grandis*
XP_010048714.1	683	0	63.543	PREDICTED: pentatricopeptide repeat containing protein At5g01110	*Eucalyptus grandis*
XP_002276556.1	617	0	76.337	PREDICTED: pentatricopeptide repeat containing protein At5g39980, chloroplastic	*Vitis vinifera*

Table D.3 continued from previous page

XP_010052727.1	422	5.26E-167	63.507	PREDICTED: sister chromatid cohesion protein DCC1 isoform X1	*Eucalyptus grandis*
XP_010060949.1	524	0	81.298	PREDICTED: probable methyltransferase PMT28	*Eucalyptus grandis*
XP_010066755.1	542	0	57.749	PREDICTED: uncharacterized protein	*Eucalyptus grandis*
XP_002263990.1	314	7.33E-152	69.108	PREDICTED: protein N-lysine methyltransferase METTL21A-like isoform X1	*Vitis vinifera*
XP_010036473.1	244	2.83E-89	69.672	PREDICTED: G patch domain-containing protein 11 isoform X1	*Eucalyptus grandis*
KHG06611.1	119	2.49E-78	94.958	tig	*Gossypium arboreum*
XP_002270988.3	719	0	76.634	PREDICTED: probable threonine tRNA ligase, cytoplasmic	*Vitis vinifera*
XP_008450440.1	637	0	61.695	PREDICTED: pentatricopeptide repeat containing protein At2g22410, mitochondrial	*Cucumis melo*

D.1 Erratum

A more recent phylogenetic calibration of *Tococa* was estimated based on a more complete species sampling within Miconieae. The results of this calibration are not included within the text as results were obtained after finalizing the whole document. The estimated tMRCA of Western and Eastern *T. guianensis* is different as the calibration of the Miconieae tribe is different. However, the relative position of both clades and the remaining Melastomataceae specimens collected during this study does not change despite the inclusion of a total of 814 sequences. Finally, a younger estimate of the tMRCA between Eastern and Western *T. guianensis* confirms that patterns of diversification between *T. guianensis* and its *Azteca* ants are a consequence of vicariance and not of codiversification.

Methods

The calibrated ITS phylogeny of *T. guianensis* specimens includes Miconieae ITS sequences available on NCBI (accession numbers listed in Table D.4), that were included in our analyses to evaluate the relative position of the *T. guianensis* specimens and reliably reconstruct the ancestral regions of *T. guianensis*. Four *Tibouchina* sequences from NCBI were included as outgroups. Using closer sequences to Miconieae as outgroups resulted in chains that did not converge. Optimal nucleotide substitution models were selected based on the Akaike Information Criterion (AIC) as implemented in **jModelTest2** (Darriba et al., 2012). A Bayesian calibration for ITS was conducted using **BEAST v1.8.4** (Drummond et al., 2012) and a birth-death model (Gernhard, 2008). The tMRCA of all Miconieae and *T. guianensis* specimens was calibrated based on the age estimations for Miconieae in Berger et al. (2016). This calibration is much more robust than the calibration based on (Morley and Dick, 2003) because is based on 10 fossils distributed across the phylogeny of Myrtales. Morley and Dick (2003) uses derived substitution rates. A normally distributed prior was placed on the Miconieae crown node with a mean of 19 million years (My) and a standard deviation of 5 My. An uncorrelated relaxed clock model was set to a prior exponential distribution with mean 0.001 and standard deviation of 0.33.

Results

Overall, no topological differences resulted from the new calibration compared to the calibration used in Chapter D; however, because the calibration of the Miconieae group based on Berger et al. (2016) is much more recent than the calibration based on Morley and Dick (2003), every node has a younger age than previously estimated. The ITS calibration of Miconieae presented in this section estimates the split of the Western *T. guianensis* at 4.39 Mya (95% Higher Probability Distribution, HPD=2.62-6.21, **c.** in Figure D.3). This western group is nested within a *T. guianensis* clade mainly containing Eastern *T. guianensis* that splits from its sister clade at 6.7 Mya (95% HPD=4.56-9.11, **b.** in Figure D.3). The tMRCA of the clade grouping both Western and Eastern *T. guianensis* is estimated at 9.68 Mya (95% HPD=6.20-13.77). As in the calibration presented in Chapter D, eastern specimens are found within the Western *T. guianensis* clade and *vice versa*.

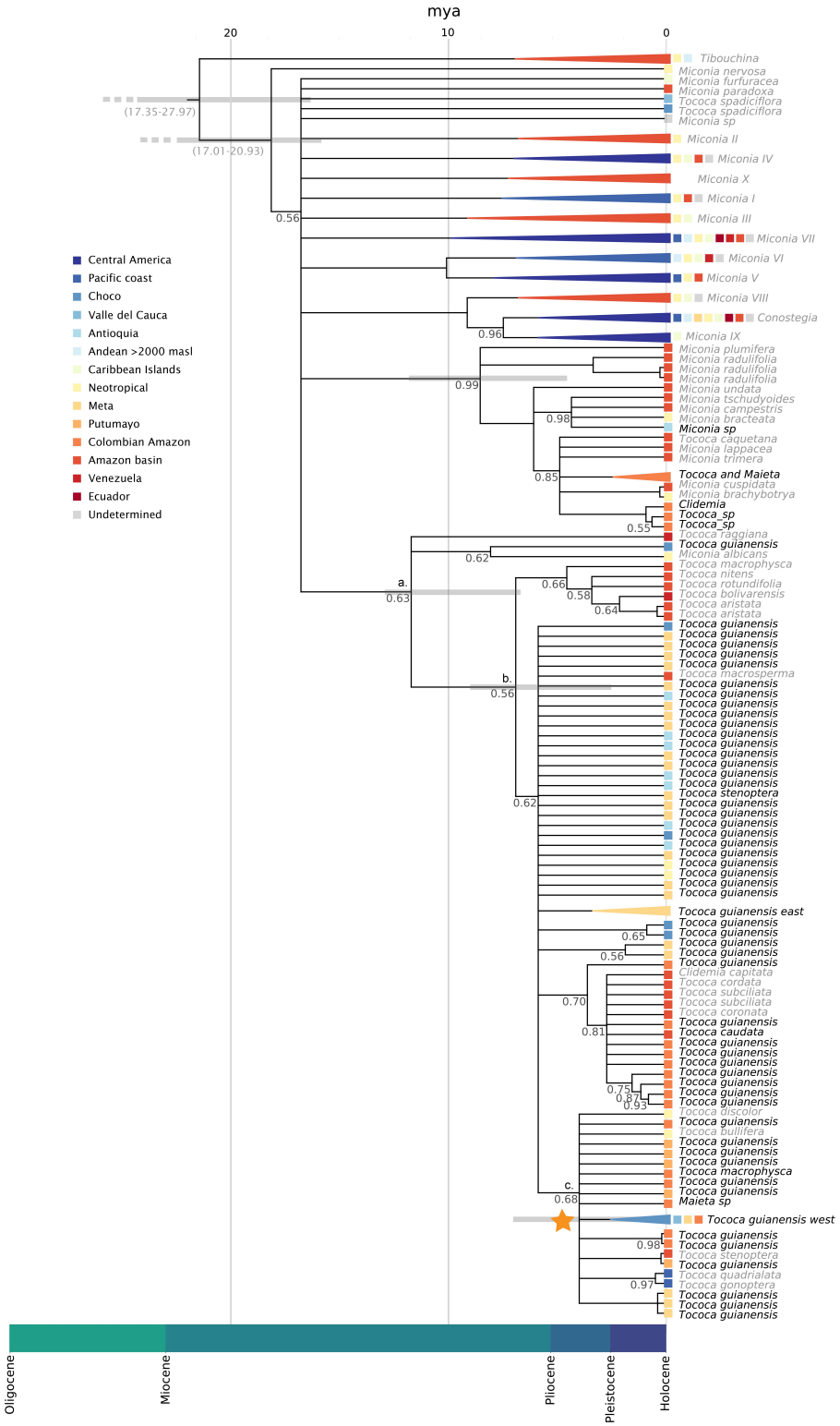


Figure D.3: ITS calibration of the Miconieae tribe and the *Tococa* specimens. Posterior probabilities between 0.5 and 0.99 are shown. All branches with lower posterior probabilities are collapsed. Colors correspond to the distribution of the species: red colors are east to the Eastern Andean Cordillera while blue colors are west. Black names correspond to specimens collected during this study and grey names correspond to NCBI reference sequences. The star marks Eastern *T. guianensis*.

Discussion

A more reliable calibration of the tribe Miconieae estimates the tMRCA to the Western *T. guianensis* clade more recently than previous calibrations. However, the patterns of nestedness of the Western clade are congruent with previous results in which Western and Eastern clades are not monophyletic but that gene flow across the Andes occurred at least until very recently. However, the east-to-west geographic structure is observed and is congruent with the geographic structure observed in *Azteca*. For both taxa, the Eastern Cordillera acts as a barrier while the Central and Western Cordillera do not. This calibrated phylogeny also confirms that the Western *T. guianensis* is more likely derived from an eastern population, rather than from a widely distributed population divided by vicariance. However, that hypothesis needs appropriate testing.

Despite congruence in the geographic structure shown by plants and ants, the time does not coincide. While *Azteca* divergence occurred around 10-12 Mya, the divergence of Western *T. guianensis* is now estimated to have occurred more recently, at around 4 Mya.

Table D.4: GenBank accession numbers and species of the reference sequences included in the ITS Miconieae calibrated phylogeny.

GenBank accession no.	Species	GenBank accession no.	Species
KM893614	*Clidemia allenii*	GQ139308	*Miconia kollmannii*
AY460549	*Clidemia capitata*	KF821631	*Miconia kraenzlinii*
KM893590	*Clidemia ecuadorensis*	KF821632	*Miconia kriegeriana*
KM893641	*Clidemia foreroi*	KJ933991	*Miconia krugiana*
KM893617	*Clidemia frate*	EF418892	*Miconia krugii*
KM893574	*Clidemia fulva*	KF821633	*Miconia labiakiana*
KM893610	*Clidemia globuliflora*	AY460514	*Miconia lacera*
KM893604	*Clidemia hammelii*	AY460515	*Miconia laevigata*
EF418891	*Clidemia japurensis*	EU069392	*Miconia laevigata*
KM893637	*Clidemia laxiflora*	KF821634	*Miconia lanceolata*
KM893567	*Clidemia mortoniana*	EF418893	*Miconia lappacea*
KM893581	*Clidemia ombrophila*	KF821635	*Miconia lappacea*
KM893606	*Clidemia petiolaris*	EU055790	*Miconia latecrenata*
KM893639	*Clidemia pittieri*	KF821636	*Miconia lateriflora*
KM893634	*Clidemia spectabilis*	EF208214	*Miconia latifolia*
KM893576	*Clidemia subpeltata*	KX073163	*Miconia lehmannii*
KF821471	*Conostegia affinis*	EU055791	*Miconia lenticellata*
KM893596	*Conostegia bernoulliana*	KF821637	*Miconia lenticellata*
AY460485	*Conostegia bigibbosa*	EU055792	*Miconia lepidota*
KM893587	*Conostegia bigibbosa*	KF821638	*Miconia leucocarpa*
KM893603	*Conostegia bigibbosa*	EF418894	*Miconia ligulata*
KM893580	*Conostegia bracteata*	KM893597	*Miconia ligulata*
KM893594	*Conostegia brenesii*	EU055793	*Miconia ligustrina*
KM893608	*Conostegia caelestis*	EU055794	*Miconia ligustroides*
EU055677	*Conostegia centronioides*	KJ418735	*Miconia lima*

Table D.4 continued from previous page

GenBank accession no.	Species	GenBank accession no.	Species
KM893613	*Conostegia centronioides*	KJ933992	*Miconia lima*
KM893643	*Conostegia cinnamomea*	KJ933993	*Miconia limoides*
KM893619	*Conostegia cuatrecasii*	KF821639	*Miconia livida*
KM893642	*Conostegia cuatrecasii*	EU055795	*Miconia lonchophylla*
AY460486	*Conostegia icosandra*	KF821640	*Miconia lonchophylla*
KM893595	*Conostegia icosandra*	KF821641	*Miconia longibracteata*
KM893618	*Conostegia icosandra*	EU055796	*Miconia longicuspis*
KF821472	*Conostegia lasiopoda*	EF418895	*Miconia longifolia*
KM975936	*Conostegia lasiopoda*	EU055797	*Miconia longispicata*
EF418809	*Conostegia macrantha*	EU055798	*Miconia loreyoides*
KM893583	*Conostegia macrantha*	KF821642	*Miconia lourteigiana*
AY460487	*Conostegia micrantha*	KF821643	*Miconia lugonis*
KM893599	*Conostegia micrantha*	EU055799	*Miconia luteola*
AY460488	*Conostegia montana*	KF821644	*Miconia lutescens*
KM893565	*Conostegia montana*	EU055800	*Miconia lymanii*
KM893582	*Conostegia montana*	KJ933994	*Miconia macayana*
KM893586	*Conostegia montana*	AY460516	*Miconia macrodon*
KM893589	*Conostegia montana*	KF821645	*Miconia macrodon*
KM893602	*Conostegia montana*	KF821646	*Miconia macrothyrsa*
KM893612	*Conostegia montana*	EF418896	*Miconia magdalenae*
KM893621	*Conostegia montana*	EU055801	*Miconia manicata*
KM893623	*Conostegia montana*	EF418897	*Miconia marginata*
KM893624	*Conostegia montana*	KF821647	*Miconia marginata*
EF418810	*Conostegia montelegreana*	KF821648	*Miconia maroana*
KF821473	*Conostegia oerstediana*	KF821649	*Miconia matthaei*
KM893575	*Conostegia oerstediana*	KF821650	*Miconia mazanana*
KM893579	*Conostegia oerstediana*	EU055802	*Miconia melanotricha*
KM893592	*Conostegia oerstediana*	EF418898	*Miconia melinonis*
KM893615	*Conostegia oerstediana*	KF821651	*Miconia mendoncaei*
KM975937	*Conostegia oerstediana*	EU055803	*Miconia meridensis*
EU055678	*Conostegia pittieri*	KF821652	*Miconia meridensis*
KM893632	*Conostegia pittieri*	KF821653	*Miconia meridensis*
KM893569	*Conostegia polyandra*	EU055804	*Miconia mesmeana*
KM893570	*Conostegia procera*	KF821654	*Miconia mexicana*
KM893571	*Conostegia pyxidata*	KM893601	*Miconia mexicana*
EU055679	*Conostegia rhodopetala*	KF821655	*Miconia michelangeliana*
KM893573	*Conostegia rubiginosa*	AY460517	*Miconia minutiflora*
AY460489	*Conostegia rufescens*	EU055805	*Miconia minutiflora*
KM893568	*Conostegia rufescens*	KF821656	*Miconia minutiflora*
KM893578	*Conostegia rufescens*	AY460518	*Miconia mirabilis*
KM893584	*Conostegia rufescens*	EU055806	*Miconia mirabilis*
KM893620	*Conostegia rufescens*	KF821657	*Miconia moensis*

Table D.4 continued from previous page

GenBank accession no.	Species	GenBank accession no.	Species
KM893633	*Conostegia setifera*	KF821658	*Miconia molybdea*
EU055680	*Conostegia setosa*	KX073164	*Miconia mulleola*
KM893622	*Conostegia sp*	KF821659	*Miconia multinervia*
AY460490	*Conostegia speciosa*	EU055807	*Miconia multiplinervia*
EU055681	*Conostegia subcrustulata*	EU055808	*Miconia multispicata*
KF821474	*Conostegia superba*	KF821660	*Miconia myriantha*
KM893577	*Conostegia superba*	KX073165	*Miconia myrtillifolia*
KM893591	*Conostegia superba*	KF821661	*Miconia nambyquarae*
KM893616	*Conostegia superba*	KJ933995	*Miconia nanophylla*
AY460491	*Conostegia tenuifolia*	KJ933996	*Miconia navifolia*
KM893566	*Conostegia volcanalis*	KF821662	*Miconia neei*
EU055682	*Conostegia xalapensis*	KJ361774	*Miconia neomicrantha*
KM893572	*Conostegia xalapensis*	KF821663	*Miconia nervosa*
KM893605	*Conostegia xalapensis*	EU055809	*Miconia nitidissima*
KF821546	*Miconia abbreviata*	KJ933997	*Miconia norlindii*
EF418877	*Miconia acuminata*	KX073166	*Miconia notabilis*
KF821548	*Miconia acuminata*	KF821664	*Miconia nubicola*
KX073142	*Miconia acuminifera*	KF821665	*Miconia nutans*
AY460501	*Miconia aeruginosa*	KF821666	*Miconia nystroemii*
EF418879	*Miconia affinis*	KF821667	*Miconia obscura*
KF821550	*Miconia affinis*	KF821668	*Miconia obtusa*
KF821551	*Miconia affinis*	EU055810	*Miconia octopetala*
KF821552	*Miconia aggregata*	KF821669	*Miconia oinochrophylla*
EU055713	*Miconia alainii*	EF418899	*Miconia oldemanii*
KF821553	*Miconia alata*	KM893585	*Miconia oligocephala*
KX073143	*Miconia albertii*	EU055811	*Miconia onaensis*
EF418880	*Miconia albicans*	KF821670	*Miconia oraria*
KF821554	*Miconia albicans*	KM893593	*Miconia osaensis*
KX073144	*Miconia alborosea*	KJ933998	*Miconia ossaeifolia*
EU055714	*Miconia alborufescens*	KF821750	*Miconia ottoschmidtii*
EF418881	*Miconia aliquantula*	KJ933999	*Miconia ottoschmidtii*
KJ933976	*Miconia alloeotricha*	AY460519	*Miconia pachyphylla*
KF821555	*Miconia alternans*	KF821671	*Miconia paleacea*
KF821556	*Miconia alternifolia*	EU055812	*Miconia papillosa*
EU055746	*Miconia amilcariana*	KF821672	*Miconia paradoxa*
KF821557	*Miconia ampla*	KJ934000	*Miconia paralimoides*
EU055715	*Miconia amplinodis*	KF821673	*Miconia paucidens*
KX073137	*Miconia anchicayensis*	KF821674	*Miconia pausana*
KF821558	*Miconia andersonii*	KJ418738	*Miconia pedunculata*
KX073145	*Miconia andreana*	KJ934001	*Miconia pedunculata*
KF821559	*Miconia androsaemifolia*	KM893611	*Miconia peltata*
KF821560	*Miconia angelana*	KM893630	*Miconia pendula*
KJ933977	*Miconia apiculata*	KF821675	*Miconia penduliflora*
AY460502	*Miconia aplostachya*	EU055813	*Miconia penningtonii*

Table D.4 continued from previous page

GenBank accession no.	Species	GenBank accession no.	Species
KX073146	*Miconia aponeura*	EU055814	*Miconia pepericarpa*
EU055716	*Miconia appendiculata*	KF821676	*Miconia pepericarpa*
KF821561	*Miconia araguensis*	EU055815	*Miconia petropolitana*
EU055717	*Miconia arboricola*	KF821677	*Miconia phanerostila*
AY460503	*Miconia argentea*	KF821678	*Miconia pileata*
KF821749	*Miconia argentimuricata*	KF821679	*Miconia pittieri*
KJ933978	*Miconia argentimuricata*	KX073167	*Miconia plethorica*
KF821562	*Miconia argyraea*	KF821680	*Miconia plukenetii*
EF418882	*Miconia argyrophylla*	EU055816	*Miconia plumifera*
EU055718	*Miconia asclepiadea*	EU055817	*Miconia plumosa*
EU055709	*Miconia aspergillaris*	EU055818	*Miconia poeppigii*
KF821732	*Miconia asperifolia*	KF821681	*Miconia polita*
KJ933979	*Miconia asperifolia*	EU055819	*Miconia polyandra*
KX073147	*Miconia asperrima*	KJ934002	*Miconia polychaete*
EU055719	*Miconia astroplocama*	KJ934003	*Miconia polychaete*
KF821563	*Miconia aulocalyx*	KF821682	*Miconia polygama*
KF821564	*Miconia aurea*	AY460520	*Miconia prasina*
KF821565	*Miconia aurea*	KX073168	*Miconia prasinifolia*
KF821566	*Miconia aureoides*	KF821683	*Miconia prietoi*
KF821567	*Miconia aymardii*	AY460521	*Miconia procumbens*
EU055720	*Miconia bangii*	EF418900	*Miconia pseudoaplostachya*
EF418883	*Miconia baracoensis*	KX073169	*Miconia pseudoradula*
EU055721	*Miconia barbeyana*	KF821684	*Miconia pseudorigida*
KF821568	*Miconia barbinervis*	KF821685	*Miconia pterocaulon*
KJ149271	*Miconia barkeri*	DQ644131	*Miconia pteroclada*
EU055722	*Miconia benthamiana*	EU055820	*Miconia pteroclada*
KF821569	*Miconia biacuta*	KF821686	*Miconia puberula*
KX073148	*Miconia biappendiculata*	EF418901	*Miconia pubipetala*
KJ933980	*Miconia bicolor*	AY460522	*Miconia pulvinata*
EU055723	*Miconia biglandulosa*	EU055821	*Miconia punctata*
KF821570	*Miconia bilopezii*	KF821687	*Miconia punctibullata*
EU055724	*Miconia biperulifera*	EU055822	*Miconia pusilliflora*
KJ418736	*Miconia blancheana*	KF821688	*Miconia pustulata*
KF821571	*Miconia boliviensis*	EF418902	*Miconia pyramidalis*
EU055725	*Miconia brachybotrya*	AY460523	*Miconia pyrifolia*
KX073149	*Miconia brachycalyx*	KF821689	*Miconia quadrangularis*
KX073150	*Miconia brachygyna*	KF821690	*Miconia quadrialata*
EF418884	*Miconia bracteata*	EU055823	*Miconia racemosa*
EU055726	*Miconia bracteolata*	EF418878	*Miconia radulifolia*
EU055727	*Miconia brasiliensis*	EF418903	*Miconia radulifolia*
GQ139307	*Miconia brasiliensis*	KF821691	*Miconia radulifolia*
EU055728	*Miconia brenesii*	EU055824	*Miconia ramboi*

Table D.4 continued from previous page

GenBank accession no.	Species	GenBank accession no.	Species
EU055729	*Miconia brevitheca*	EU055825	*Miconia reducens*
EU055730	*Miconia brunnea*	KF821692	*Miconia regelii*
EU055731	*Miconia bubalina*	AY460524	*Miconia resimoides*
EU055732	*Miconia buddlejoides*	KF821693	*Miconia rigida*
EU055733	*Miconia bullata*	EU055826	*Miconia rigidiuscula*
KX073151	*Miconia buxifolia*	EU055827	*Miconia rimalis*
EU055734	*Miconia cabucu*	EU055828	*Miconia robinsoniana*
EU055735	*Miconia caesariata*	KF821694	*Miconia robusta*
EU055710	*Miconia calignosa*	KF821695	*Miconia rosea*
KF821572	*Miconia calocoma*	KF821696	*Miconia rubens*
KM893629	*Miconia calocoma*	AY460525	*Miconia rubiginosa*
EU055736	*Miconia calvescens*	KX073170	*Miconia rubricans*
EU055737	*Miconia calycina*	KJ934004	*Miconia rubrisetulosa*
KJ933981	*Miconia calycopteris*	KJ934005	*Miconia rubrisetulosa*
KF821573	*Miconia campestris*	EU055829	*Miconia rufa*
KF821574	*Miconia capitellata*	AY460526	*Miconia rufescens*
EU055738	*Miconia capixaba*	EU055830	*Miconia rugosa*
EU055739	*Miconia carnea*	KF821697	*Miconia saldanhaei*
KF821575	*Miconia castaneiflora*	KF821698	*Miconia salebrosa*
KX073152	*Miconia cataractae*	EU055831	*Miconia salicifolia*
KF821576	*Miconia caudata*	KX073171	*Miconia salicifolia*
EU055740	*Miconia caudigera*	EU055832	*Miconia samanensis*
KF821627	*Miconia caudigera*	AY460527	*Miconia sancti-philippi*
AY460504	*Miconia centrodesma*	EU055833	*Miconia schlechtendalii*
KF821577	*Miconia centrodesma*	EU055834	*Miconia schlimii*
KM893609	*Miconia centrosperma*	KM893625	*Miconia schlimii*
EF418885	*Miconia ceramicarpa*	AY460528	*Miconia schnellii*
EU055741	*Miconia ceramicarpa*	KF821699	*Miconia schunkei*
EU055742	*Miconia ceramicarpa*	EU055835	*Miconia sclerophylla*
KF821578	*Miconia ceramicarpa*	EF418904	*Miconia selleana*
EU055743	*Miconia cerasiflora*	EU055836	*Miconia sellowiana*
EU055744	*Miconia cerasiflora*	EU055837	*Miconia septentrionalis*
KM893627	*Miconia cerasiflora*	AY460535	*Miconia serrulata*
EU055745	*Miconia cercophora*	KF821700	*Miconia sessilifolia*
KF821579	*Miconia cernua*	KX073172	*Miconia silverstonei*
EU055748	*Miconia chamissois*	EU055838	*Miconia simplex*
EU055749	*Miconia chartacea*	AY460529	*Miconia sintenisii*
DQ644129	*Miconia chionophila*	EU055839	*Miconia skeaniana*
KX073153	*Miconia chrysocoma*	EU055840	*Miconia smaragdina*
EU055750	*Miconia chrysophylla*	KX073173	*Miconia smithii*
KF821580	*Miconia ciliata*	EF418876	*Miconia sp*
KF821581	*Miconia ciliata*	FJ358430	*Miconia sp*
EU055751	*Miconia cinerascens*	FJ358431	*Miconia sp*
EU055752	*Miconia cinerascens*	KF821702	*Miconia sp*
KJ418737	*Miconia cinerea*	KF821704	*Miconia sp*

Table D.4 continued from previous page

GenBank accession no.	Species	GenBank accession no.	Species
EU055753	*Miconia cinnamomifolia*	KJ361766	*Miconia sp*
KF821582	*Miconia cipoensis*	KJ361767	*Miconia sp*
KF821583	*Miconia cladonia*	KJ361769	*Miconia sp*
KX073154	*Miconia clypeata*	KJ361771	*Miconia sp*
KF821584	*Miconia cocoensis*	KJ361772	*Miconia sp*
EU055754	*Miconia collatata*	KJ361773	*Miconia sp*
KM893635	*Miconia colliculosa*	KJ361775	*Miconia sp*
KF821585	*Miconia commutata*	KJ933975	*Miconia sp*
EU055755	*Miconia concinna*	KJ933983	*Miconia sp*
KF821586	*Miconia confertiflora*	KJ933985	*Miconia sp*
KF821587	*Miconia corallina*	KJ933986	*Miconia sp*
KF821588	*Miconia coriacea*	KJ934006	*Miconia sp*
KX073156	*Miconia coronata*	KJ934008	*Miconia sp*
EU055756	*Miconia corymbiformis*	KM893600	*Miconia sp*
EU055757	*Miconia costaricensis*	KM893607	*Miconia sp*
EU055758	*Miconia crassinervia*	KM893626	*Miconia sp*
KF821589	*Miconia cremadena*	EU055841	*Miconia sphagnicola*
EU055759	*Miconia crocata*	KX073174	*Miconia spicellata*
DQ644130	*Miconia crocea*	AY460530	*Miconia spinulosa*
EU055760	*Miconia crocea*	KF821701	*Miconia splendens*
EU055761	*Miconia cubatanensis*	KF821705	*Miconia squamulosa*
EU055762	*Miconia cubensis*	KF821706	*Miconia stelligera*
EU055747	*Miconia cuprea*	EU055842	*Miconia stenobotrys*
KJ933982	*Miconia curvipila*	KF821707	*Miconia stenophylla*
EF418886	*Miconia cuspidata*	EU055843	*Miconia stenostachya*
KF821590	*Miconia cuspidatissima*	KF821708	*Miconia stevensiana*
KF821591	*Miconia cyanocarpa*	KX073175	*Miconia stipitata*
KF821592	*Miconia cyathanthera*	EU055844	*Miconia striata*
EU055763	*Miconia dapsiliflora*	KF821709	*Miconia suaveolens*
KF821593	*Miconia decurrens*	EU055845	*Miconia subcompressa*
EU055764	*Miconia delicatula*	KJ149273	*Miconia subcompressa*
EU055765	*Miconia denticulata*	KJ149274	*Miconia subcompressa*
KF821594	*Miconia desmantha*	KJ149276	*Miconia subcompressa*
EF418887	*Miconia desportesii*	KJ149277	*Miconia subcompressa*
KF821595	*Miconia diaphanea*	KJ149278	*Miconia subcompressa*
KF821596	*Miconia diegogomezii*	KJ149279	*Miconia subcompressa*
KM495209	*Miconia diegogomezii*	KJ149283	*Miconia subcompressa*
EU055766	*Miconia dielsiana*	KJ149284	*Miconia subcompressa*
EU055767	*Miconia discolor*	KJ149285	*Miconia subcompressa*
KF821597	*Miconia dispar*	AY460531	*Miconia subcorymbosa*
EU055768	*Miconia dissita*	KX073176	*Miconia summa*
KF821598	*Miconia dissitiflora*	EU055846	*Miconia superba*
KM893598	*Miconia dissitiflora*	EU055847	*Miconia sylvatica*
KF821599	*Miconia dissitinervia*	KF821710	*Miconia tabayensis*
KM893636	*Miconia dissitinervia*	KX073177	*Miconia tamana*

Table D.4 continued from previous page

GenBank accession no.	Species	GenBank accession no.	Species
AY460506	*Miconia dodecandra*	KF821711	*Miconia tentaculifera*
EU055769	*Miconia dodecandra*	KF821712	*Miconia tetragona*
EU055770	*Miconia dodecandra*	EU055848	*Miconia tetrandra*
FJ358429	*Miconia dodecandra*	AY460532	*Miconia tetrastoma*
KF821600	*Miconia dodecandra*	AY460533	*Miconia theaezans*
KM495208	*Miconia dodecandra*	EU055849	*Miconia theizans*
EF418888	*Miconia dolichopoda*	KF821713	*Miconia theizans*
KF821601	*Miconia dolichopoda*	KF821714	*Miconia theizans*
KX073157	*Miconia dolichopoda*	EU055850	*Miconia thomasiana*
KF821602	*Miconia dolichorrhyncha*	KF821715	*Miconia tiliifolia*
AY460507	*Miconia donaeana*	KF821716	*Miconia tillettii*
KY782466	*Miconia donaeana*	KF821717	*Miconia tinifolia*
EU055771	*Miconia doriana*	EF418905	*Miconia tomentosa*
KF821603	*Miconia dorsaliporosa*	EF418906	*Miconia tonduzii*
KF821604	*Miconia dorsiloba*	KF821718	*Miconia tonduzii*
AY460508	*Miconia duckei*	KF821719	*Miconia tonduzii*
KX073158	*Miconia dunstervillei*	KF821720	*Miconia tonduzii*
KF821605	*Miconia egensis*	KX073178	*Miconia tovarensis*
KX073159	*Miconia elaeoides*	KF821721	*Miconia traillii*
EU055772	*Miconia elata*	EU055851	*Miconia trianae*
KF821606	*Miconia elata*	EU055852	*Miconia triangularis*
KF821607	*Miconia elegans*	EF418907	*Miconia trimera*
KJ933984	*Miconia ellipsoidea*	EU055853	*Miconia trinervia*
EU055773	*Miconia elvirae*	KF821722	*Miconia trinervia*
KF821608	*Miconia eriocalyx*	EU055854	*Miconia triplinervis*
KF821547	*Miconia erioclada*	EU055855	*Miconia tristis*
KX073160	*Miconia erioclada*	EF418908	*Miconia tschudyoides*
KF821609	*Miconia eriodonta*	AY460534	*Miconia tuberculata*
AY460509	*Miconia ernstii*	KF821723	*Miconia tuberculata*
EF418889	*Miconia ernstii*	KX073179	*Miconia tuberculata*
KX073161	*Miconia erosa*	DQ644132	*Miconia turquinensis*
EU055774	*Miconia fasciculata*	EU055856	*Miconia turquinensis*
KF821610	*Miconia ferruginata*	AY460536	*Miconia ulmarioides*
AY460510	*Miconia ferruginea*	EF418909	*Miconia undata*
KF821611	*Miconia ferruginea*	KF821724	*Miconia uninervis*
KJ149282	*Miconia ferruginea*	KM893638	*Miconia uninervis*
EU055775	*Miconia floribunda*	KX073180	*Miconia urticoides*
KX073162	*Miconia floribunda*	KF821725	*Miconia valerioana*
AY460511	*Miconia foveolata*	EU055857	*Miconia valtherii*
KF821549	*Miconia fragilis*	KX073181	*Miconia velutina*
EU055776	*Miconia friedmaniorum*	KF821726	*Miconia victorinii*
KM893628	*Miconia friedmaniorum*	EU055712	*Miconia villonacensis*
KF821612	*Miconia fuertesii*	EF418910	*Miconia viscidula*
EU055777	*Miconia furfuracea*	KF821727	*Miconia wagneri*
KF821613	*Miconia galeiformis*	KF821728	*Miconia walterjuddii*

Table D.4 continued from previous page

GenBank accession no.	Species	GenBank accession no.	Species
KF821614	*Miconia glandulifera*	EU055858	*Miconia willdenowii*
EU055711	*Miconia glutinosa*	KF821729	*Miconia wilsonii*
KF821615	*Miconia glutinosa*	KJ934007	*Miconia woodsii*
KF821616	*Miconia glutinosa*	KJ149275	*Miconia xenotricha*
KF821617	*Miconia gonioclada*	KJ149280	*Miconia xenotricha*
EU055778	*Miconia goniostigma*	KJ149281	*Miconia xenotricha*
KF821618	*Miconia gracilis*	KF463039	*Tibouchina brevisepala*
EU055779	*Miconia grandidentata*	KF463044	*Tibouchina dimorphophylla*
KF821619	*Miconia grandifoliata*	JQ730208	*Tibouchina martiusiana*
KJ933987	*Miconia granulata*	JQ730221	*Tibouchina pereirae*
EF418890	*Miconia gratissima*	KF463048	*Tibouchina saxosa*
KF821620	*Miconia gratissima*	JQ730229	*Tibouchina sp*
KM893631	*Miconia grayumii*	KF821773	*Tococa aristata*
KF821621	*Miconia heliotropoides*	MF785442	*Tococa aristata*
EU055780	*Miconia hemenostigma*	AY460547	*Tococa bolivarensis*
KF821622	*Miconia hirtella*	AY460548	*Tococa broadwayi*
KJ933988	*Miconia hispidula*	KF821774	*Tococa bullifera*
KF821623	*Miconia holosericea*	AY460550	*Tococa caquetana*
KF821624	*Miconia hondurensis*	KF821775	*Tococa carolensis*
EU055781	*Miconia hookeriana*	AY460551	*Tococa caudata*
KJ933989	*Miconia hottensis*	KF821776	*Tococa cordata*
AY460512	*Miconia howardiana*	AY460552	*Tococa coronata*
EU055782	*Miconia hyemalis*	EU055895	*Tococa discolor*
EU055783	*Miconia hymenonervia*	AY460553	*Tococa gonoptera*
KJ149272	*Miconia hypiodes*	AY460554	*Tococa guianensis*
EU055784	*Miconia hypoleuca*	MF785330	*Tococa guianensis*
EU055785	*Miconia ibaguensis*	AY460555	*Tococa macrophysca*
AY460513	*Miconia impetiolaris*	AY460556	*Tococa macrosperma*
KJ418734	*Miconia inaequipetiolata*	JN032012	*Tococa macrosperma*
EU055786	*Miconia inconspicua*	KF821703	*Tococa macrosperma*
KX073139	*Miconia indicoviolacea*	AY460557	*Tococa nitens*
KF821625	*Miconia intricata*	AY460558	*Tococa perclara*
KF821626	*Miconia ioneura*	EU055896	*Tococa platyphylla*
EU055787	*Miconia jahnii*	EF418922	*Tococa quadrialata*
KF821628	*Miconia jahnii*	AY460559	*Tococa raggiana*
KJ933990	*Miconia jashaferi*	AY460560	*Tococa rotundifolia*
EU055788	*Miconia javorkaeana*	KF821778	*Tococa sp*
KF821629	*Miconia javorkaeana*	EU055897	*Tococa spadiciflora*
KX073155	*Miconia javorkaeana*	KF821779	*Tococa stenoptera*
KM893640	*Miconia jefensis*	AY460505	*Tococa subciliata*
KM893644	*Miconia jefensis*	AY460561	*Tococa subciliata*
KF821630	*Miconia jorgensenii*	MF785407	*Tococa subciliata*
EU055789	*Miconia jucunda*	KF821777	*Tococa symphyandra*
KF821742	*Miconia karlkrugii*		

APPENDIX

E

APPENDIX DISCUSSION

E.1 Lineage specificity between *Tococa* and *Azteca*

Methods

DNA barcoding in Chapters 2 and 4 show evidence of two monophyletic and well differentiated *Azteca* lineages (Western and Eastern). However, the differentiation between Eastern and Western *T. guianensis* lineages are less clear and neither of the lineages is reciprocally monophyletic. Thus, two interaction matrices were built from the species specificity analysis. The lineage-level matrix separates different Western and Eastern *T. guianensis* lineages regardless of their non-monophyly. The species-level matrix considers Western and Eastern *T. guianensis* lineages as a single species (the *T. guianensis clade* in Figure 4.5 in Chapter 4 or **a.**, Figure D.3 in Appendix D). No sequences of *T. guianensis* plants from the Santander populations were successfully obtained, for that reason, they are kept as an individual lineage in the lineage-level matrix. In the species-level, those samples are merged within *T. guianensis* because specimens were fertile and could have been identified as such. The same was done for some *T. guianensis* specimens from Antioquia (*T. guianensis* Ant. in Figure 5.3, Chapter 5). Other specimens from other species from which no sequences were successfully obtained were added to the analysis but represented as independent lineages in Figure 5.2 in Chapter 5.

The species specificity coefficient is calculated for each node in the plant and ant levels of the matrix and it goes from zero to one, meaning low and high species specificity (Poisot et al., 2012). The coefficient was calculated between the *Azteca* and *Tococa* lineages for both matrices including all specimens collected during this study and the interactions reported for the ant reference sequences. The matrices were analyzed using the *specieslevel* function in **Bipartite v.2.08** package for **R v.3.4.0** (Dormann et al., 2008). In both cases, the observed species specificity values of plants and ants were compared to values estimated from null matrices. Those matrices were calculated using the *vaznull* function in **Bipartite v.2.08**, which randomizes the interactions, ensures that every item has at least one interaction and keeps the same dimensions as the observed matrix. Interactions recorded for the *Azteca* sequences from NCBI were retrieved from their original publications and included in the matrix, with the purpose of highlighting that the genus *Azteca* associates to more than one plant species. The bipartite figure was generated using **Bipartite v.2.08** and the trees added by hand, both sets of figures including other ant and plant species collected during this study.

Results and discussion

Network representations of the ant-plant associations show that *Azteca* lineages collected from *T. guianensis* are also present in *T. cordata*, *T. bullifera* and *T. macrophysca* specimes. In one case, a specimen of Western *Azteca* is associated to a Eastern *T. guianensis* that was collected to the west of the Eastern Cordillera. Similarly, *T. guianensis* is host of other genera of ants different from *Azteca*, incuding *Pheidole*, *Crematogaster*, *Myrmelachista* and *Solenopsis*. However, *Tococa*-associated *Azteca* represents lineages independent from those associated with *Cecropia* and *Cordia*. Opposite to *Azteca*, *Pheidole* associates with a higher diversity of host genera, including not only *Tococa* species but also *Clidemia*, *Conostegia*, *Miconia* and *Ossaea* (Figure 5.3 in Chapter 5).

The mean species specificity coefficient is higher in ants than in plants (0.87 *versus* 0.77 in Table E.1) indicating that ants tend to be more specific. Consequently, individual species specificity coefficients of *Azteca* lineages indicate higher specificity compared to those in *Pheidole* in both the species and the lineage-level matrices (Table E.2). Within *T. guianensis* lineages, the least specific lineage is *T. guianensis* east1 as it associates with two different *Azteca* Western and Eastern lineages and with *Solenopsis* and *Tapinoma* ants (Figure 5.3 in Chapter 5). The remaining *T. guianensis* lineages have higher species specificity coefficients. Other plants have a coefficient of one that is likely to be an artefact of the reduced sampling and the no inclusion of all possible ant associates (as the study is not the focus on those). Similarly, rarely collected ants that represent opportunistic species have a coefficient close to one because only one interaction of all possible ones is represented in the matrices. Finally, coefficients calculated from the observed data does not vary from those calculated from null, randomized matrices (Figure E.1).

The same patterns are observed when comparing the phylogenies of the main ant and plant lineages. Lineages within the wide *T. guianensis* clade, including the *T. stenoptera, T. macrophysca, T. cordata* and *T. bullifera* species, are associated mainly with the *Tococa-Azteca* monophyletic clade (Figure 5.2 in Chapter 5). These results are observed at the lineage and the species-level networks (Figure 5.3 in Chapter 5), indicating that *Azteca* lineages are specific to the host genus, that other opportunist ants can inhabit *Tococa* plants and that *Pheidole* as a genus appears to be more generalist. However, the analysis does not discern between *Pheidole* lineages and that can bury any genus-specificity patterns. Finally, the sampling here presented does not include other non-Melastomataceae host plants that can potentially host different or the same lineages of *Azteca*, therefore the results of this work can be partial and a more complete survey of ant-plants and their colonies is needed at all localities.

Table E.1: Summary statistics for the species specificity coefficient calculated for the species and lineage-level matrices incorporating both and only plant and ant levels of the matrix.

Species specificity	**Species-level matrix**		**Lineage-level matrix**	
	Observed	**Null model**	**Observed**	**Null model**
Both levels				
Mean	0.83	0.78	0.78	0.74
Median	1.00	0.69	1.00	0.69
Standard Deviation	0.26	0.23	0.26	0.23
p-value		0.19		0.23
Plant level				
Mean	0.77	0.71	0.75	0.72
Median	1.00	0.69	0.69	0.69
Standard Deviation	0.29	0.25	0.26	0.23
p-value		0.23		0.26
Ant level				
Mean	0.87	0.72	0.81	0.72
Median	1.00	0.69	1.00	0.69
Standard Deviation	0.23	0.25	0.26	0.23
p-value		0.21		0.32

Table E.2: Species specificity coefficient estimated for each node of the species and lineage level matrices. A coefficient close to zero indicates low species specificity and a coefficient close to one indicates high species specificity.

Species-level matrix			
Plant	**Coefficient**	**Ant**	**Coefficient**
Tococa guianensis	0.21	*Pheidole*	0.17
Cordia	0.36	*Azteca east2*	0.54
Tococa cordata	0.41	*Azteca sp2*	0.54
Cecropia	0.47	*Tapinoma*	0.54
Maieta poeppigii	0.55	*Allomerus*	0.68
Tococa bullifera	0.55	*Crematogaster*	0.68
Tococa caquetana	0.69	*Myrmelachista*	0.68
Clidemia	1	*Azteca east1*	1
Conostegia sp	1	*Azteca east3*	1
Henriettella cuneata	1	*Azteca sa1*	1
Miconia	1	*Azteca sa2*	1
Ossaea bullifera	1	*Azteca west6*	1
Tococa macrophysca	1	*Azteca-C1*	1
Tococa sp.	1	*Azteca-C2*	1
Tococa spadiciflora	1	*Azteca-Cecropia1*	1
Tococa stenoptera	1	*Azteca-Cecropia2*	1
		Azteca beltii	1
		Azteca forelii	1
		Azteca nigricans	1
		Azteca ovaticeps	1
		Azteca pittieri	1
		Azteca sp1	1
		Camponotus	1
		Nylanderia	1
		Solenopsis	1
Lineage-level matrix			
Plant	**Coefficient**	**Ant**	**Coefficient**
T. guianensis east1	0.29	*Pheidole*	0.16
Cordia	0.36	*Azteca east2*	0.36
Tococa cordata	0.41	*Tapinoma*	0.41
Cecropia	0.47	*Azteca sp2*	0.47
T. guianensis an	0.47	*Azteca west6*	0.55
T. guianensis east2	0.47	*Myrmelachista*	0.55
Maieta poeppigii	0.55	*Azteca east1*	0.69
T. guianensis west	0.55	*Allomerus*	0.69
Tococa bullifera	0.55	*Crematogaster*	0.69
T. guianensis east 3	0.69	*Solenopsis*	0.69
T. guianensis east4	0.69	*Azteca east3*	1
T. guianensis sa	0.69	*Azteca sa1*	1
Tococa caquetana	0.69	*Azteca sa2*	1
Clidemia	1	*Azteca-C1*	1
Conostegia sp.	1	*Azteca-C2*	1
Henriettella cuneata	1	*Azteca-Cecropia1*	1

Table E.2 continued from previous page

Species-level matrix			
Miconia	1	*Azteca-Cecropia2*	1
Ossaea bullifera	1	*Azteca beltii*	1
T. guianensis west1	1	*Azteca forelii*	1
T. guianensis west2	1	*Azteca nigricans*	1
Tococa macrophysca	1	*Azteca ovaticeps*	1
Tococa sp.	1	*Azteca pittieri*	1
Tococa spadiciflora	1	*Azteca sp1*	1
Tococa stenoptera	1	*Camponotus*	1
		Nylanderia	1

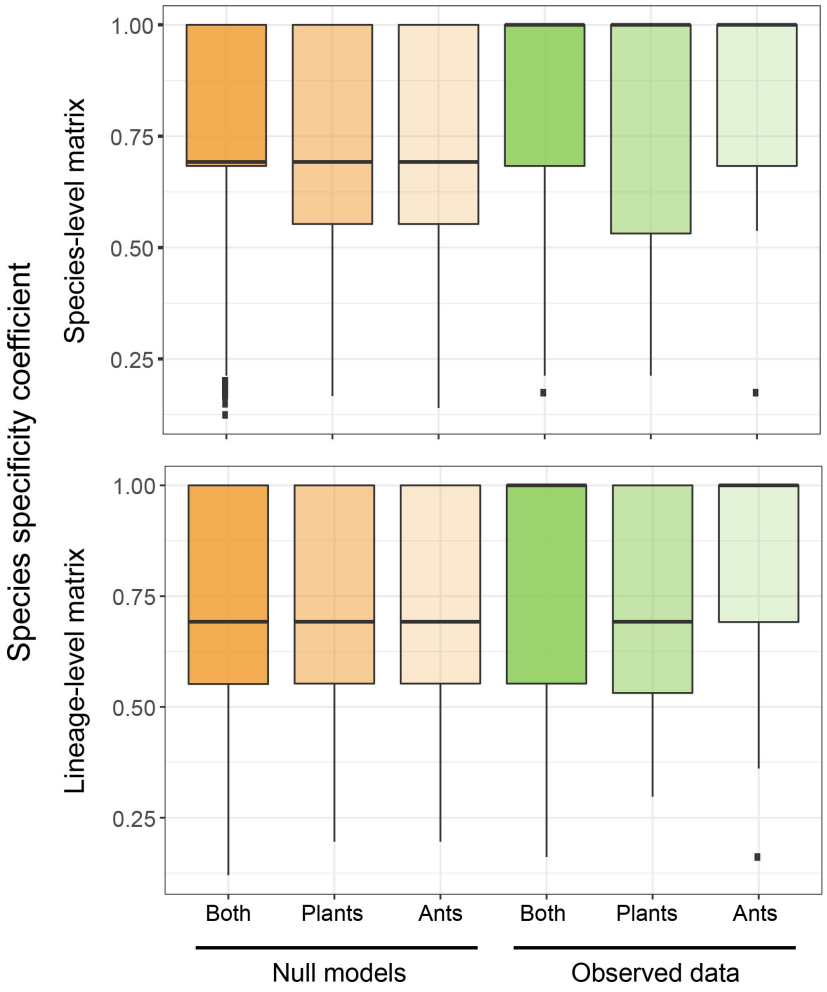


Figure E.1: Distribution of species specificity coefficient for both, plant and ant levels calculated from the observed interactions (green) and from null randomized matrices (orange) with the same dimensions of the observed matrix. Values were calculated for the species and the lineage-level matrices.